DOCUMENT RESUME

ED 085 421 TM 003 366

AUTHOR	Dragositz, Anna, Ed.
TITLE	Curriculum Innovations and Evaluation: Proceedings of the Association for Supervision and Curriculum Development Pre-Conference Seminar (Princeton, March 8-9, 1968).
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Association for Supervision and Curriculum Development, Washington, D.C.
PUB DATE	69
NOTE	103p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Behavioral Objectives; *Curriculum Development; *Curriculum Evaluation; *Evaluation Techniques; Innovation; Inservice Programs; Program Evaluation; *Program Improvement; Research Design; Test Construction; Testing; Ungraded Primary Programs

ABSTRACT

This report on curriculum innovation and evaluation contains several sections which deal with procedures for the development of measurement instruments; the uses and limitations of tests; research design and the interpretation of results; a definition of objectives; and the role of evaluation in curriculum innovation. Also included are three examples of curriculum evaluation projects, specifically illustrations of the National Longitudinal Study of Mathematical Ability; in-service teaching programs; and an ungraded primary school. (NE)

ED 085421

TM 003 366

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FILMED FROM BEST AVAILABLE COPY

CURRICULUM INNOVATIONS AND EVALUATION

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Dorothy Urban

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

Proceedings of the Association for Supervision and Curriculum Development Pre-Conference Seminar

Educational Testing Service
Princeton, New Jersey

ED 085421

CURRICULUM INNOVATIONS AND EVALUATION

Proceedings of the
Association for Supervision and
Curriculum Development
Pre-Conference Seminar

Held on March 8 and 9, 1968 at
Educational Testing Service
Princeton, N. J.

Seminar Chairman: John S. Helmick
Proceedings Editor: Anna Dragositz

TABLE OF CONTENTS

Page

Introduction..... 1
John S. Helmick

The Role of Evaluation in Curriculum Innovation..... 3
Henry S. Dyer

Definition of Objectives..... 21
S. Donald Melville

Research Design and the Interpretation of Results..... 30
George E. Temp

Uses and Limitations of Tests..... 42
Miriam M. Bryan

Development of Measurement Instruments: Procedures and Problems..... 57
Thomas F. Donlon

Illustrations of Curriculum Evaluation Projects

The Role of Large Scale Projects in Curriculum Evaluation with Examples From the National Longitudinal Study of Mathematical Abilities..... 73
Leonard S. Cahen

In-Service Education Programs..... 85
Thomas S. Barrows

The Ungraded Primary School..... 92
J. Robert Cleary

Concluding Remarks..... 99
John S. Helmick

INTRODUCTION

John S. Helmick
Vice President
Educational Testing Service

I am very pleased to have the opportunity to welcome you to this preconference seminar. It is satisfying to know that so many people (nearly fifty from about twenty states) are interested in what we feel is a topic of critical importance.

It may help to put into perspective what we are trying to do here today if we note some of the changes that have taken place over the last ten or fifteen years in education and in Educational Testing Service.

Fifteen years ago, ETS essentially supplied tests and testing services. While we did provide supporting research and advice, we were basically test-oriented, with emphasis on the printed paper-and-pencil test. Education at that time seemed fairly stable. We were all confident that we would be doing tomorrow just about what we did yesterday.

Now, change is the order of the day. It is good to innovate, and I have some faith that in many cases, change for the sake of change may have a positive effect, if the teachers and staff are fully involved in producing the change. It is true, however, that innovation may be simply a glorified or continuous Hawthorne effect, and change may not be improvement--particularly when we automatically adopt somebody else's innovation. We need to know what we are doing and how it affects our results. Fortunately, there has been increasing professional interest in determining the relation between the changes we are making in curriculum and outcomes in terms of student behavior.

It is fortunate that the profession is becoming active because the public is now very much interested in what is happening to the funds that are being allocated to education. In Princeton, for example, possibly for the first time in its recorded history, the school budget has just been twice defeated in public referendum. I think perhaps the general public has been applying cost-benefit analysis to education with-

out knowing the terminology. In effect, they have been asking: What are we getting for what we are paying? We, as professionals, have a clear obligation to provide better answers to that question than we have in the past.

Measurement and evaluation are very much involved in the answer to such questions. I say "measurement" and not "testing" because we have gone beyond the point of pulling a published test off the shelf, giving it at the annual testing time, and saying we've done our evaluation. Measurement should be an integral part of education--the systematic collection of information to aid in the decision-making process. This is the role that ETS is now attempting to fill.

This seminar cannot make experts out of participants in a day and a half. We do hope, however, that all of us can gain a little understanding. From our vantage point, we hope that we can communicate some things to you, and I am sure you will be able to communicate things to us, from your varied experiences with the day to day problems of the schools, that will help us all see the problems and the ways of dealing with them more clearly.

ASCD Pre-Conference Seminar at
Educational Testing Service
March 8-9, 1968

THE ROLE OF EVALUATION IN CURRICULUM INNOVATION

Henry S. Dyer
Vice President
Educational Testing Service

The main theme of this talk is that curriculum evaluation is full of bewildering problems that are not to be solved by simple, pat formulas that yield simple, pat answers. This fact should not be surprising to anybody who gives the matter a modicum of thought. Educational evaluation has to be complicated and difficult because education itself is, without a doubt, the most complicated and difficult business with which man, in his inadequate wisdom, has ever tried to contend. It is also, without a doubt, his most important business--if he expects to survive to the year 2000. If education is to go anywhere at all in this time of rapid change and impossible dilemmas, it is incumbent on all of us, all the time, to keep on trying to figure out what it's doing for children now, what it ought to be doing, and how to bring current practice a little closer to current hopes. This figuring-out process is what I mean by evaluation in all its multitude of forms. We're not very good at it yet. The answers we get are usually pretty hazy. The methods we employ are full of uncertainty and short on rigor. But unless we keep trying, the chances are that this great big amorphous enterprise we call education will not take us anywhere at all. In what follows, I'll try to clarify the picture of curriculum evaluation by giving a few examples that may help to identify the several kinds or levels of curriculum evaluation that are needed to get us out of our current confusions.

I. Curricular Evaluation in a Narrow Frame

One way we have stumbled upon to keep our confusions at bay is to think of curriculum as a collection of bits and pieces, and to content ourselves with assessing the effects of each curricular bit one at a time. Let me start off with an example of this approach from my own experience in trying to evaluate a piece of curriculum some twenty years ago. The piece in question was the required freshman course in English

composition at Harvard--English A of sacred memory. I suppose nobody ever had an opportunity such as I had at that time to design and conduct an evaluation study that could be more exactly what your simon-pure evaluator might want an evaluation study to be. Everything was working in my favor.

This course had a long tradition. Time out of mind, it had been required of all freshmen entering Harvard except those who could prove by examination that they were such good writers as to be beyond help. The number of such exceptions was very small--perhaps one or two per cent of the freshman class each year.

The reputation of English A varied according to the mood of the faculty, the mood of the students, and who happened to be saddled with the job of running the course and the army of section hands who taught it. In the last years of its existence, the professor in charge was a remarkable man--a scholar, a poet, a novelist, and withal a dedicated teacher of freshman composition. It has always been a mystery to me how he ever let himself get conned into the management of English A, with all its headaches and all the brickbats that were perpetually being thrown at it. Nevertheless, there he was doing it, and working hard to make it a good and effective course.

Came along, now, another professor, who also was a not inconsiderable figure in the literary world, and a dedicated teacher, who thought English A was ripe for extinction, and who argued that he could put on a different course that would do better in one semester what English A did in two. Furthermore, he thought that the army of section men should be demobilized and that his course could be taught just as well in large lecture sections--300 or 400 students per section--in which films, film clips, and batches of short exercises would turn the trick of teaching freshmen how to write. The grading of themes would be minimized. He was so sure of his vision of pedagogical truth that he requested, and got, the opportunity to teach such an experimental course. Let the faculty assign 300 freshmen to his course, rather than English A, and he would prove his point.

By fortunate coincidence, these two professors had a congenial relationship with each other. So they invited me to lunch one day, while the plans for the new course were still brewing, and asked me whether I would whomp up some way of "scientifically" evaluating their rival modes of going at the job of teaching freshmen to write. I would have carte

blanche in designing the experiment and the measures to be used in assessing the relative effectiveness of the two courses, and they would cooperate to the hilt. In short, the set-up was the sort of thing your earnest evaluator dreams about, but rarely gets a chance to put into practice.

The plan we worked out was this: I would assign the incoming freshmen on a random basis--some to the old course, and some to the new. The arrangement was that, working with me, and with each other, the two course directors would deve. writing exercise to be given at the start of both courses and again at the end of the first semester. My crew would administer the exercise, and I would lay out the procedures by which it would be graded and the results analyzed.

All this was done. After the two sets of papers had been written (i.e., the pretest and the posttest), we coded them for identification purposes, removed the names and dates, and arranged them in batches of twenty papers each, thus:

		Time written	
		September	January
Course taken	Old	a b c d e s s s s s	a b c d e j j j j j
	New	a'b'c'd'e' s s s s s	a'b'c'd'e' j j j j j

The symbol $\underline{a}_s$ stands for a paper written in September by a student assigned to the old course; $\underline{a}_j$ stands for a paper written in January by the same student; $\underline{a}'_s$ stands for a paper written in September by a student assigned to the new course; and $\underline{a}'_j$ stands for a paper written in January by that same student; etc. The papers in each batch were sufficiently scrambled so that the reader could not figure out who wrote which one, in which course, or when. Each batch of twenty papers was ranked from one to twenty by two readers working independently, and the average of the ranks assigned to any paper was the score that paper received. The

question we asked of the papers in each batch was whether the old course or the new course showed the greater upward shift in ranks between September and January.

When we aggregated all the data, we got some interesting, but not very cheerful, results. First, it became pretty clear that the new course was no better or worse than the old course in developing writing ability as measured by the exercise that the two professors agreed should be an adequate measure of it. Second--and this upset the whole faculty no end--it was equally clear that in neither course were the January papers any better than the September papers. So, naturally, the decision was made to go with the new course as the cheaper way to achieve nothing much.

Although the results of this study disappointed everybody, nevertheless, by the canons of evaluation, it was a very neat study that produced data which were unassailable. All of which shows how the evaluation process can work when conditions are right. It also shows why rigorous evaluation can be shattering to educators.

Now I want to turn to a quite different example of curriculum evaluation, still in a narrow frame, but illustrative of somewhat different approaches to evaluation. During the time when the Physical Science Study Committee was constructing the PSSC physics course at MIT, ETS undertook the evaluation of the several elements of the course as it was being put together and revised. I shall come back to that aspect of the evaluative process in a moment, but first I want to tell you about the experience of one of the creators of the course--a physicist, Professor Walter Michels of Bryn Mawr. Professor Michels felt that the test and questionnaire results we were getting as the course was being developed were not suffcient indicators of how good the course actually was. To see whether the course was really paying off, he wanted to conduct a follow-up study in which the college performance of students who had taken PSSC in high school would be compared to the college performance of students who had taken traditional physics in high school. He theorized that if the PSSC course was doing to students what it was supposed to do, more PSSC students than non-PSSC students would be signing up for physics in college. Furthermore, he figured that, since PSSC was a

"process-of-discovery" course designed to teach students to think the way scientists think, then the PSSC students ought to do better in college physics, would have a more mature attitude toward science, and would more frequently major in the subject. In short, even though the new physics was only a one-year course, he felt that, if it was as good as he and others claimed, it should have superior long-term effect. This search for the ultimate pay-off is what, in the new jargon, evaluators are now calling "summative evaluation." Walter Michels felt he couldn't be comfortable until he had some summative evaluation of PSSC physics.

Well, he and I got together and spent a morning discussing this approach, and I spent most of the time trying to discourage him from undertaking the impossible. I kept telling him, over and over, that the kind of information he was looking for was simply out of reach. Since the PSSC kids and their conventional counterparts would be scattering to colleges all over the country, we could not expect that they would be exposed to college physics courses that could be compared in any way. Furthermore, we didn't have any dependable information about the content or quality of the physics courses the students had taken in high school. Since teachers differ, courses differ, even though they may carry the same labels and use the same instructional materials. It would indeed be reasonable to suppose that there are as many different physics courses as there are teachers teaching physics, whether labeled PSSC or something else. Finally, even if you could assume some reasonable degree of uniformity in the "treatments" each kind of student got in high school and in college, it would be too much to expect that any college physics teacher--present company excepted of course!--could supply any valid information about the mental characteristics of his students, about their attitudes toward science, or whether they think the way scientists think. Who ever heard of a college teacher who can report reliably on the thought processes that pass through the brains of his students? The notion is absurd on the face of it.

But Professor Michels refused to be discouraged by my doubts. He went ahead and organized his follow-up study, complete with attitude scales, rating scales, and the like. A year later he was back in my office confessing total defeat. The mish-mash of incomplete data he

had managed to accumulate added up to exactly no information at all. As a first-class scientist, he was able to face the fact and admit that the whole effort had been a bust.

This episode is not to be taken to mean that curriculum evaluation founded on follow-up studies is forever impossible, or that it is undesirable. I bring the matter up only to emphasize that the difficulties in getting interpretable data from such studies are more enormous than most people imagine, and that accordingly one needs to consider whether the agony involved in going after long-term follow-up data is likely to be anywhere near commensurate with their usefulness.

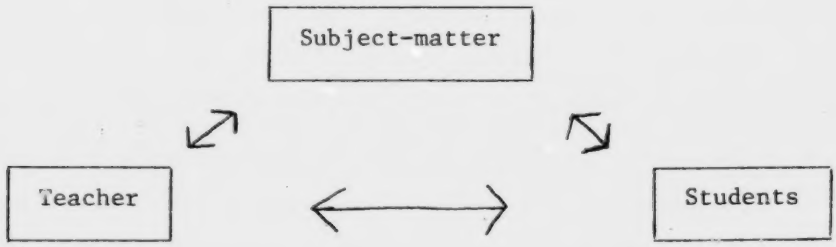
I have already mentioned that we were working on another kind of evaluation in connection with PSSC. Its purpose was to get periodic feedback to the curriculum makers as the course was being developed and tried out. This sort of activity has latterly been labeled "formative evaluation," to suggest what needs to be done while a curriculum is in its formative stages. In this case, it consisted of developing and giving a whole series of unit tests to check up on how the different parts of the course were actually working in the classrooms where the new material was being tried out. In addition, batches of questionnaires were inflicted on the participating teachers from time to time, to see what they thought of the stuff and to get their views on how their students were responding to it. All of this was supplemented by classroom visits by some of the curriculum constructors, who travelled around the country to see what was going on out there where the action was.

For obvious reasons, this so-called formative evaluation is a very important kind. It is not, of course, experimental in any formal sense; it can't tell you much about the ultimate pay-off; and it is, in fact, purely descriptive. But it is absolutely vital as a means of finding out in detail how the new material is working, what kind of stuff is working for what kinds of students, and what changes need to be made to make it work better. As you try to fashion the individual components of a new course, you desperately need to know, as you go along, how they are bouncing off the minds of students to see what is connecting and what is not. You don't worry about experimental designs, control groups, and tests of statistical significance; you do worry about the adequacy of the

week-to-week and month-to-month feedback, so that as you hone the course down into its final shape you will have some assurance that it will do the job you intend it to do.

One of the important incidental values we discovered in connection with the PSSC experience--and one that has informed subsequent efforts in curriculum construction--is that when you oblige the curriculum makers to create tests to measure the effects of their materials as they go along, you induce in them a much clearer and more specific set of ideas about their curricular objectives. We've found that probably the one best way in all the world to get a person to put his objectives into so-called "behavioral terms" (awful expression!) is to force him to cast those objectives in test questions aimed at eliciting the kinds of student responses he claims he is trying to help them learn. So one might almost say that the exercise of careful test making is a prerequisite to careful curriculum making.

When you get into this sort of formative evaluation of a developing curriculum, you realize that the curriculum is not just a set of materials and rules of procedure. It has in it three interacting elements: the subject-matter, the teacher, and the students, like this:



And you have to be concerned about all the elements at once, and all the interactions among them, if you are to get any idea of how successful your curriculum innovations are turning out to be.

II. Broadening the Frame

This triangular model of the curricular process suggests, I hope, how complicated any adequate evaluation is bound to be. Even so, I don't think it is complicated enough to illustrate what we are really up against. And this is because it rests on a still too narrow conception of the term "curriculum." Over the last 10 or 15 years or so--since the coming of the new math, the new science, the new social studies, and the

new everything else--we have tended to think of curriculum almost as if it were no more than a particular course or sequence of courses arranged around a particular body of subject matter. In so doing we have, I think, tended to overemphasize the needs of the academic disciplines as they exist in the minds of university scholars, and have underemphasized the needs of the students as they are trying to learn to understand themselves and to cope with an increasingly confusing and threatening world. I suspect that we have become somewhat too concerned with the pure and particular cognitive outcomes of individual courses and have neglected to look at other associated outcomes--both cognitive and noncognitive--of the whole school experience. Which is to say that, in our efforts to evaluate--whether they be summative or formative, descriptive or experimental, formal or informal--we have paid far too little attention to the kinds of side effects that studying any particular course may have. Somehow or other, we have to take the whole educational picture into account and try to see what the unplanned, indirect effects may be on students when they grapple with any particular chunk of subject matter that we may happen to think may be good for them. For instance, we need to be asking whether a new course may be reducing the students' interest in other subjects in the curriculum, or whether, indeed, it may be reducing or killing off their enthusiasm for the particular new course we're trying to teach. For example, how many kids have been turned away permanently from all math after struggling to understand the new math?

We don't have any good answers to questions like this, and I suspect the reason we don't arises from our present tendency to hold off ambiguity by thinking only of bits and pieces. In the old days, back in the 1920's, the concept of curriculum innovation was quite different from what it now seems to be. The model then, you may remember, was Jesse Newlon's Denver Program of Curriculum Revision which attracted imitators all over the country. The theory behind that program was that the person who should be at the center of the curriculum revision process was the classroom teacher, not the university scholar, as is the case today. To overstate the contrast: in those days, the university scholars were called in as consultants at the pleasure of the teachers; nowadays, the teachers are called in as consultants at the pleasure of the scholars.

Furthermore, the emphasis in the Denver Revision was on the total school program and the pupil's total career through the educational system, rather than on a single course in a single year as was the case with the PSSC. Newlon's point was that the teacher must not only be the central agent in curriculum change, but also that the new curriculum must emerge from the students' needs, as the teachers were able to observe those needs through direct contact with the students and the community--i.e., their personal needs, life needs, social needs, career needs, as well as their purely intellectual needs. This goal was never really attained, but I think the goal is still worthy, even though it seems to have been largely forgotten.

Please do not mistake my meaning. I do not mean to be anti-intellectual or anti-scholarly. I do not doubt for a moment that if the range and quality of the intellectual life of the country is to meet the needs of the times, we must have people who are going to be excited by the intellectual challenges in the academic disciplines so that they will pursue them and advance them. We also must be concerned with developing a public that can understand and appreciate what is going on in the several disciplines, so that the big social decisions of the future will be informed decisions. In short, we must not, in this day and age, underestimate the importance of scholarly endeavor to the life of our times, and hence to organized education. Nevertheless, as things are going now, it seems to me that too few people are paying attention to the social contexts in which the intellectual disciplines must operate. There's too little attention to priorities. Scarcely anybody seems to be thinking about the criteria for determining, for instance, what kinds of physics, or social studies, or math, or art, or whatever, are best for <u>these</u> particular kids in <u>this</u> particular school--in the ghetto, the rural backwater, the affluent suburb.

How, then, does evaluation fit into the broader approach to curriculum development? How are we to go about assessing the totality of the effects of our educational programs in all their complex variety? The answers are not going to be easy. We are unlikely to find them by the simplistic process of seeing how the student products of such pro-

grams make out in college--as the old Eight-Year Study tried to do. That was certainly one of the really great early efforts in curriculum reform, and its evaluation was in the hands of one of the really great innovators in evaluation--Ralph Tyler. It has therefore always been a puzzle to me why the directors of the Eight-Year Study felt it incumbent on them to rest their case on the quality of the college performance of the students who went through the program. One can state the matter categorically: success in college is simply not an adequate test of whether kids are getting what they need out of pre-college programs of any kind.

So we simply have to reckon with bewilderment, and try to find our way out of it as best we can. Our bewilderment indeed is exacerbated these days by a number of new curriculum developments that have recently come on the educational scene. Let's look at some of them and consider some of the tough evaluation problems they raise.

Many of you are probably familiar with the Oak Leaf Project that is going on outside of Pittsburgh. Known as IPI (Individually Prescribed Instruction), it represents a tremendous effort--a really serious approach to the individualization of instruction that we have been advocating for years, but not doing much about it.

IPI's strength is that it concerns itself with the diagnosis of student learning needs; it therefore tries to take into account the differences in developmental processes of the children, their speed of absorption, and the ways in which they learn best. It attempts to keep a running account of each pupil's accomplishments--his specific strengths and weaknesses--and to organize all this material in such a way that the teacher can prescribe the kinds of exercises that are likely to be of most benefit to the pupil in moving him onward and upward.

This provision of constant feedback to pupils and teachers alike is what one might call "continuous evaluation." Such evaluation is a never-ending process in which we evaluate not the course as a whole nor the educational program as a whole, but the progress of the individual student as he goes along. IPI has made some important advances in making such a system operational, but it is still confronted with vast problems--not the least of which is the inadequacy of the diagnostic measures that

are needed to make the flow of information about all aspects of pupil performance as comprehensive as it really needs to be. We know, for instance, that there are important individual differences in learning styles, but the measures available for detecting those differences are still primitive. We know that attitudinal processes have a lot to do with what children learn and how they learn it, but our measures of such things are still so crude that they scarcely serve the purpose of the sort of continuous evaluation that IPI requires, if it is to fulfill its promise as an approach to individualized instruction.

The consequence is that as matters now stand, the experimenters with IPI have had to narrow their concerns pretty largely to the development of basic cognitive skills, and individualization has come to mean not much more than adapting instruction to the rate--and only the rate--at which the child learns. In brief, the IPI experience to date demonstrates the general principle that the teaching-learning process is inevitably hamstrung by shortcomings in the instruments and techniques available for continous evaluation.

Another similar type of curriculum innovation, presenting similar problems for the evaluator, is to be found in the Nova Schools in Florida. The central element in the Nova system is what are known as learning activity packages (LAPS). In this scheme, the classroom teachers are again being brought back into the center of the curriculum development picture. They review all the new curricular materials as they come on the market and choose from among them the exercises, activities, projects, readings, etc., that they consider most appropriate for the students in their charge. The combination becomes the set of learning packages for the students. The students work at their LAPS sometimes alone, sometimes together, and sometimes with their teachers. When the students complete an agreed-upon segment, they present themselves for evaluation. Periodically, as experience with the LAPS accumulates, the packages are revised to bring them up to date and to re-tailor them to fit pupil needs. The whole process looks quite revolutionary, until you recall the old "contract method" of the Winnetka Plan that was dreamed up some 40 years ago. And one cannot help but wonder whether the Nova LAPS may not go the way of

the Winnetka Plan, when the initial excitement dies down. The best insurance against such an eventuality is, I submit, some hard-headed evaluation built into the system itself to bring its operators up short, when and if it begins to slip.

Finally, let's think a moment about the so-called systems approach to curriculum innovation. One example of it is Project PLAN, now in process of development by the American Institutes for Research (AIR). Project PLAN starts off by defining all the objectives that anyone might need to reach to become a functioning member of society. From these ultimate objectives, one then derives by logical processes the whole sweep of intermediate objectives that have to be reached in succession as the pupil moves up through the curriculum from preschool to a job. It is a breath-taking approach to the total educational process, with its own evaluation presumably built in at every step of the way. Evaluation, in this case, consists in ascertaining that each intermediate step in the system leads logically and inexorably to the ultimate goals. All one has to do is to make sure that the system is as logically consistent as it appears to be.

Another program that is taking the total systems approach to curriculum innovation is one that has come out of the Office of Education. It is called ES'70, i.e., Educational System for the Seventies. Here, too, the emphasis is on making academic work continuously relevant to the kinds of vocational and citizenship demands that will be made upon youngsters when they get out into the world, and to make obvious to them while they are still in school what the demands will be. Project PLAN and ES'70 have a common ancestor that was well-known back in the 1920's, but has since been largely forgotten--namely, the old social utility theory of curriculum construction. Remember that? Remember how we used to do a job analysis of societal requirements in terms of frequency and cruciality in order to determine what kids should be taught from grade one and up? The social utility theory died somewhere around 1940, but it now seems to have slipped right back into life. Nobody has noticed the miracle, probably because we now have a new name for it--RELEVANCE! This reincarnation of social utility is no doubt all to the good, but

over-reliance on it as the single principle of curriculum development can again be its undoing, for two reasons: first, society is changing so fast that we can't be sure of what demands will be placed on the present-day kindergartener twenty years from now, and second, a curriculum that is solely concerned with helping individuals adapt to the needs of society leaves unattended the educational problem of helping them discover the means for creating a Good society.

III. The All-Encompassing Frame

The common element in these most recent curriculum developments--and one that I think is hopeful, if it can be maintained--is the thrust toward a broader concept of curriculum. But we are hardly all the way there yet. The vast majority of people, inside and outside of the schools, still think of a curriculum almost exclusively in terms of small pieces of more or less traditional subject matter--algebra, American history, chemistry, English composition, etc.--rather than in terms of the totality of experiences intended to affect the growth of pupils in all its many and interrelated dimensions--physical and mental, emotional and attitudinal, social and personal. I think that until we can get most school people to think habitually about their work in these grander terms, all of our strenuous efforts at curricular reform, and the evaluation thereof, are likely to carry us nowhere. People simply have got to get used to the idea that any alteration in any part of a school program--i.e., in the materials and methods of instruction, the administrative arrangements, the rules and regulations, the training of teachers, the contacts with parents and community agencies--produces multiple alterations in all the other parts which can have multiple and differential effects on pupils.

When we think of curriculum in these extremely broad terms, how shall we think of an evaluation scheme commensurate with so grand a design? This is a hard question. Let's creep up on it by considering the difference between measurement and evaluation.

Some people make the mistake of assuming that the two terms are essentially synonymous. They are not. Good evaluation always includes some sort of measurement; but measurement is a necessary, not a sufficient

ingredient of the total evaluation process. Educational measurement, broadly defined, consists in ordering individuals in accordance with their responses to test situations associated with any kind of learning--cognitive and noncognitive--that takes place under instruction. Defined this broadly, it seems obvious to me, and I hope clear to you, that the measurement process is indeed an indispensable part of the evaluative process.

But measurement isn't the whole of it by any means. Evaluation also means making value judgments about what to measure, what is important to look at in the educational scene, what it is that needs to be observed. In addition, evaluation means the whole crucial business of coming to a decision about what to do as a consequence of whatever the measures show.

Let me elaborate on these two points by referring to the Coleman Report on equality of educational opportunity, which churned up so much interest and controversy. In planning that study, thousands of judgments had to be made both about what to measure and about how to interpret whatever was measured. It is on many judgments like these that much of the controversy about the Report has centered. For instance, it was decided early in the game that one of the pupil variables that ought to be measured is what has come to be called "locus of control." Does a pupil feel that his future destiny is largely under his own control, or largely under the control of his external environment? Does he think success depends mostly on his own efforts or on luck? A measure of this was put into the study because earlier studies had suggested that locus of control might have a good deal to do with how well a pupil got along in school. And, indeed, the Coleman study found this to be the case: students who felt that they were in control of their environment, rather than vice versa, tended to get higher scores in reading and arithmetic and other academic subjects. This was especially true of disadvantaged youngsters.

But there is room for quarrel about how these results should be interpreted. In the Coleman Report, measures of academic achievement are invariably treated as the dependent variables, and measures of such things as locus of control are treated as independent variables. In a

manner of speaking, this implies that a kid's academic achievement to some extent depends on (is caused by) the degree to which he feels he has control over his environment. Therefore, according to this line of reasoning, if you want to beef up the achievement of disadvantaged children, you see to it that, in one way or another, they are given a better sense of control over their environment. This is one kind of interpretation. But one <u>can</u> interpret the measures in quite a different manner. One can say that one of the outcomes of education of prime importance is giving a child some control over his environment, and that therefore "locus of control" ought to be thought of as the dependent variable, with reading, arithmetic, etc. as the independent variables. This interpretation means you should concentrate your efforts on helping the pupil read better, so that he will be able to get better command of his environment. The way you swing on an issue of this kind can have very large consequences for the kinds of decisions you make in determining how the priorities should lie in planning the curriculum. It is these kinds of decisions that are at the heart of the evaluation process.

The reason I have spent so much time on this particular issue is that it highlights a tendency on the part of most of us to think too exclusively about curriculum in terms of cognitive outcomes--the three R's, science, social studies--and not enough in terms of the hard-to-measure attitudinal outcomes, just <u>because</u> they are hard to measure. This is unfortunate, for it prevents us from considering the kinds of trade-offs that we ought to be thinking about when we think about curriculum matters--trade-offs such as how much intellectual boredom is to be traded for how much skill in addition or subtraction.

Having said all this, and having no doubt stirred up more confusion about evaluation than existed before this speech began, let me now guarantee your complete bewilderment by giving you my nice, big, global, hazy, over-simplified definition of what total evaluation of the total curriculum means to me. Here goes:

> Evaluation is a process for reaching decisions about the total educational program and its numerous components on the basis of relevant, dependable, and interpretable information about students, the con-

ditions of their learning, and the actual events that take place in classrooms.

The operating concept in this definition is, of course, that "evaluation is a process for reaching decisions." This means it involves people--all the people inside and outside the educational system who are, or ought to be, concerned enough about it to try constantly to make rational decisions about it.

Now, in order to try to pull together all the elements that ought to enter into the total evaluation process, let me show you my favorite picture (latterly called "Dyer's Wheel" by some of my increasingly bored and less reverent colleagues).

I call it the "Student Change Model of an Educational System" (Figure 1). At the center is the educational process (EP), which consists of all the things that are done to and by students inside the school and that are intended to make a favorable difference in them as they move from time 1 (t_1) to time 2 (t_2). Times 1 and 2 bound any slice of the total program on which you care to focus--primary school, junior high, senior high--or they may encompass the whole bit from pre-kindergarten through the Ph. D. In order to get some reading on how effective the educational process is, we have to know not only what happens inside the EP box, but we also have to know as much as it is possible to know about pupil input at t_1 and pupil output at t_2--i.e., the characteristics of the kids as they enter the EP and their characteristics as they emerge from it. By "characteristics" I mean such things as health, physical fitness, knowledge, skills, hopes, aspirations, vocational competence, social attitudes, values, etc., for what we are concerned about is how all these characteristics change between t_1 and t_2.

Just knowing inputs and outputs and the events in EP is not enough, however, to tell us all we need to know about how the total curriculum, and each of its components, may be functioning. In addition, we have to know everything we possibly can about three kinds of conditions that surround the whole system: home conditions, community conditions, and school conditions. The arrows that form the spokes of the wheel are intended to suggest that, in order to understand the whole why and how

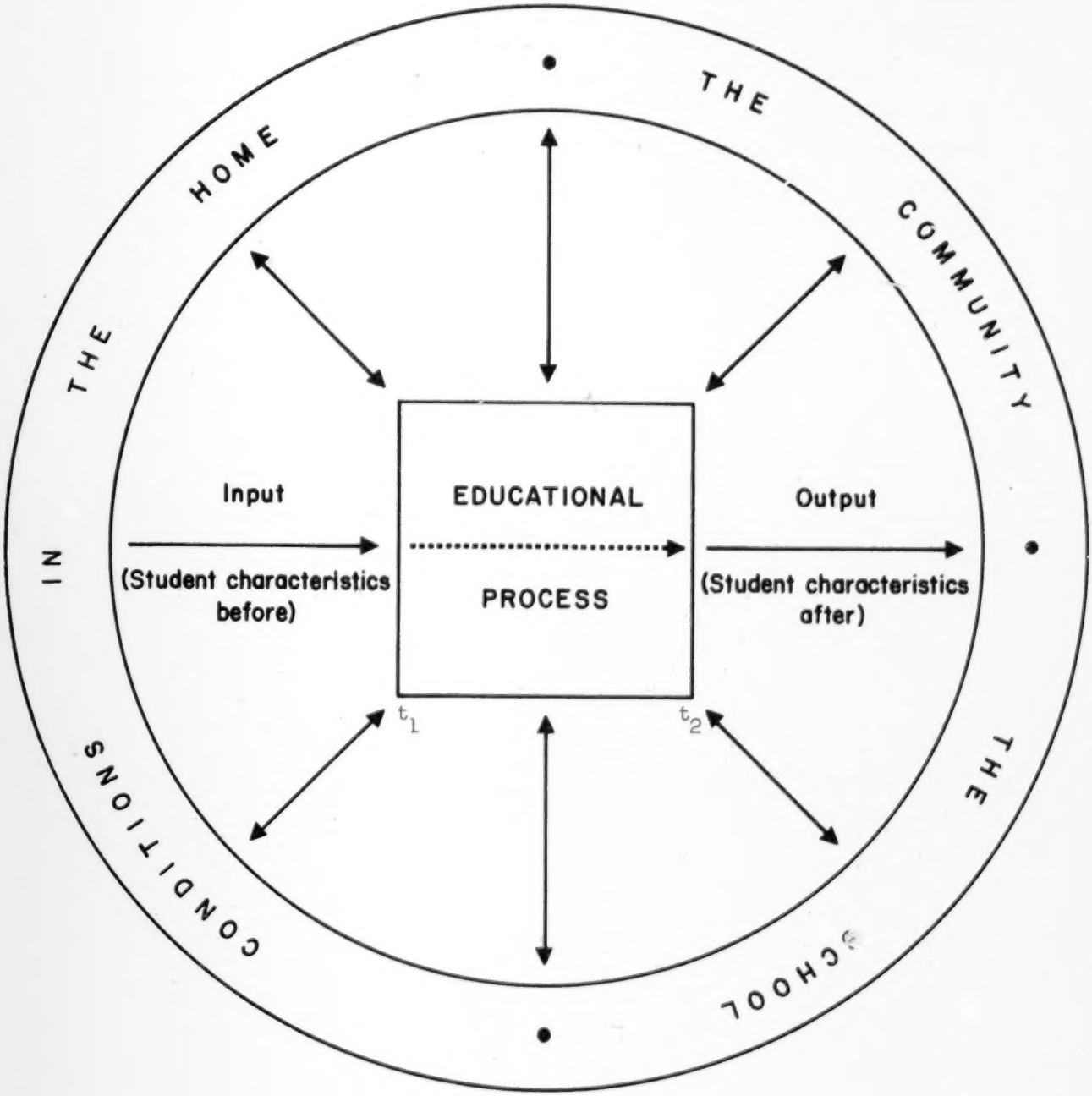


Figure 1. Student Change Model of an Educational System

of all that is going on, we must also take into account the interrelationships between the surrounding conditions and the EP variables, and the input and output variables.

All of this is meant to suggest that the total evaluation process is vastly complicated and full of tough problems. It's an ideal to be sought, not a reality that has been won. If it leaves you utterly bewildered, so much the better, for in my book, the recognition of bewilderment in these matters is the beginning of educational wisdom.

DEFINITION OF OBJECTIVES

S. D. Melville
Executive Director,
Instructional and Advisory Services
Educational Testing Service

I suppose the most widely quoted definition of evaluation in the literature today is that provided by Lee Cronbach (1962). He defined evaluation as "the collection and use of information to make decisions about an educational program." The decisions to which he referred can be classified under three broad headings: feedback, judgments, and instructional research.

Feedback refers to those situations in which the evaluation process is used by the program developer during the process of building his program. He may use the evaluation procedure periodically during the development process to determine whether or not he has been successful in achieving intermediate goals which he has set for himself. In other instances, he may want to evaluate the success of very specific pieces of his program, such as the presentation of a film or a special demonstration, to determine whether or not this activity was making the kind of contribution to his program that he intended. Jerome Bruner (1966), in his book Toward a Theory of Instruction, provides an excellent description of the way in which the process of evaluation can make a significant contribution to the development of a program. I am sure you are all aware of the fact that a recent ASCD yearbook (Wilhelms, 1967), was concerned with the use of evaluation in this fashion. I think that the chapters by Paul Diederich and Frances Link provide particularly good illustrations of how the feedback from evaluation can contribute to the development of programs in a school setting.

Presented at Association for Supervision and Curriculum Development
Pre-Conference Seminar, Princeton, New Jersey, March 8, 1968

Evaluation can also aid in the process of judging. This function of evaluation is likely to be of particular value to school administrators who are faced with the problem of trying to decide if a specific program should be instituted in their school system. During the past five years, the rate at which programs are being developed has increased markedly. As a result, school administrators have an ever-widening array of programs in a given area from which to choose. For example, look at the many different offerings that are provided in an area like mathematics or the teaching of foreign languages. As a result, the conscientious administrator, recognizing that he should not choose solely on the basis of attractiveness of packaging, cost, or the personality of the salesman, looks to the process of evaluation as a means of helping him make a sound decision. In fact, the sheer number of decisions of this sort that must be made by an administrator concerning the value of possible products and programs has led to the establishment of an organization called Educational Products Information Exchange (called EPIE for short). An excellent account of the importance of this kind of product-program evaluation has been provided by Robert Stake (1967) in the first issue of The EPIE Forum.

Instructional research is the third type of situation in which the evaluation process can make a contribution. Instructional research is concerned with answering such questions as: What aspects of the program are responsible for influencing change? What is the nature of the change which they generate? Are there differential changes among the students? Such information has importance for learning activities beyond the evaluation of the immediate program. As Cronbach (1962) points out, hopefully evaluation studies will go beyond reporting on this or that course and help us to understand educational learning. Such insight will, in the end, contribute to the development of all courses rather than just the course under test. The importance of this kind of evaluation is particularly emphasized by Hastings (1966). He provides a number of illustrations of instructional research studies that have contributed to our knowledge of the learning process.

The evaluation process, then, provides us with the means for improving the development of a specific program through feedback, the appropriateness of products or programs for a given school setting, and

increasing our knowledge of human learning by helping us discover why we obtain certain outcomes and for whom.

The basic activity in the evaluation process is that of collecting information. In the recent revival of interest in evaluation, the most striking characteristic of the new models is their emphasis on the collection of a wide variety of data. In the earlier, more primitive models of evaluation with which we are familiar, the gathering of information was frequently confined to obtaining pre- and post-test scores from alternate forms of a standardized achievement test. Considering the magnitude of the data which these studies failed to collect, it is no wonder that the literature is full of reports indicating that the experimental treatment seemed to have little, if any, effect on the learning process. In contrast, the contemporary evaluator may be charged with collecting more information than he knows what to do with.

A number of different methods have been proposed for classifying the various kinds of information that one needs to gather in the evaluation process. Christine McGuire (1967), for example, provides a fairly elaborate scheme which takes into consideration sources of information, input data, and outcomes. Scriven (1967) classifies his information in terms of educational objectives, follow-up information, and secondary effects. However, the scheme that I found most interesting and useful was that proposed by Stake (1967)--primarily, I guess, because I had little difficulty in relating the suggestions of others to his scheme.

Stake would have us collect information in what amounts to three separate time periods. The first body of information is related to what he calls the "antecedents." These include any kind of data which can be used to describe the students prior to treatment and which could conceivably bear some relationship to the success of the treatment in producing the desired outcomes. The second kind of data he labels as "transactions." These are all of the specific activities which are included in the program or instructional process. Finally, "outcomes" would include all the data relating to the results of the treatment.

The observations that are made to describe adequately the antecedents, transactions, and outcomes will be quite varied in nature. Any observational technique is fair game, providing, of course, that the

evaluator has a sound reason for using it and is capable of recognizing its limitations. For example, observations may be based upon any kind of test (objective, performance, projective, essay), informal teacher observations, anecdotal records, check lists, and so forth. As Cronbach (1962) points out, even the technique of programmed instruction can be used to provide useful information. In fact, even some of the "unobtrusive measures" described by Webb and others (1966) should not be ignored.

The object of these observations should not be confined to the students themselves, but should be extended to include anything which could have some bearing on the outcome of the treatment or which could be affected by the treatment. For example, one would want to describe in some detail specifically what kind of treatment was applied, noting particularly those instances in which elements were introduced into the program which were not originally intended. Often, too, the attitudes of teachers toward the program can influence considerably its success or failure. Certainly one would not want to ignore the effects which the program had on others in the school environment, such as non-participating students and teachers as well as school administrators. There are occasions, too, where the introduction of a program in a school produces community reactions which can have serious consequences for the program's success.

In other words, the observations which are required for an adequate evaluation program represent a monumental task for the evaluator. It is almost impossible for him to anticipate all of the observations which he ought to make. However, careful preplanning should give him some assurance that there will not be any gaping holes in his data after the program has been completed. Further, he should not hesitate to include observations during the course of the program for which he could not anticipate a need.

In the evaluation process, of equal importance to the collection of observations relating to antecedents, transactions, and outcomes is that of determining the goals, objectives, or intentions of the program developer. Presumably, the innovator had some kind of overall purpose or rationale for beginning his activity in the first place. I suppose in most instances, motivation for beginning such a project begins with dissatisfaction with what is going on at the moment. Convinced

that he can build a better mousetrap, the innovator sets about creating a general framework within which he plans to develop a new set of materials. Years ago the innovator would probably have proceeded either alone or with a few colleagues to create his program without the aid of any formal evaluation and with just his general framework to guide him. In recent years, however, program development has become a far more sophisticated activity, requiring extensive outside funding, teams of collaborators with various special backgrounds, and provision for extensive evaluation procedures.

In many instances the introduction of an evaluator into a program development project spells trouble. There are probably many reasons why evaluators tend to produce friction in their contacts with other members of a program staff. Certainly the evaluator is at a disadvantage since ordinarily he is not a subject-matter specialist in the area being investigated. Often, too, he is looked upon by the other members of the staff with suspicion since, in some sense, he represents a threat to them by being charged with judging the value of their products or techniques. Bruner (1966) reports that he tried changing the name of the evaluation group to Instructional Research Group. After a period of trial, the Instructional Research Group reported that "an evaluative branch of any organization is likely to be suspect. Even the Supreme Court is not always able to keep out of trouble."

It seems to me, however, that the major friction-producing act on the part of evaluators is their insistence upon the precise definition of goals, outcomes, or objectives. In theory at least, the more precisely one can define the objectives for a given program, the more clean-cut and precise will be its evaluation. In a nutshell, the argument runs: If you don't know exactly what you plan to do, it is difficult, if not impossible, to tell whether or not you have been successful in doing it.

To a larger extent than anyone, those who have been concerned with programmed instruction seem to be the ones mainly responsible for an increased emphasis upon precise statements of educational objectives in behavioral terms. The nature of their work is such that such specificity is absolutely essential. Anyone who is not convinced of this should read Glaser and Reynolds' (1964) description of the process of developing a programmed instruction sequence. Evaluators of other types

of programs have embraced this principle much to the dismay of the program innovators with whom they are working. They have insisted that the innovator must spell out all of the objectives which he intends to achieve in behavioral terms and do so before any work on the program is started. There is nothing quite so pitiful to watch as a heavy-handed evaluator in the act of bullying a group of program innovators into stating their objectives in behavioral terms before they really have a clear idea of what they are up to.

There are a few heretics who are doubting the necessity of spelling out objectives too specifically and too early in the development of a program. One of these is J. Myron Atkin (1968). One of the points he makes is that the possible goals of the program are so numerous that it is impossible for one to identify them with any degree of precision. He argues further that "there are important learning outcomes that cannot be anticipated when the objectives are formulated." Many of these, he claims, would be just as important, if not more so, than those which can be identified. He feels also that if you identify a body of objectives, it is quite likely that the curriculum will be restricted to cover only these objectives; some of the long-range, more important, and less readily identified objectives are likely to be lost in the shuffle. He points out that the typical curriculum developer is quite likely to start out with only a general idea of the kinds of changes which he wishes to make in the students' behavior. Some major changes become more clearly defined only as he works with the children and has a chance to see what is and is not possible for them to accomplish. The final point he makes is that the evaluator seems to assume that only those things which can be measured are worthwhile striving for. Atkin would claim that "Goals are derived from our needs and from our philosophies. They are not and should not be derived primarily from our measures." A similar set of cautions has been voiced by Eisner (1967).

Certainly it would seem that there are good arguments on both sides of this debate. It is quite true that to the extent that we are able to specify our objectives in behavioral terms, we can do a better job of evaluation. On the other hand, it seems rather foolish to restrict the creative energy necessary for the development of a good program by binding it in a strait jacket of objectives. It would seem that the

statement of objectives should be dependent upon the type of program one is developing as well as its stage of development. For example, I can see very well how a very precise statement of objectives would be necessary for the creation of a programmed instruction project, but it seems to me that other types of curriculum programs can begin with rather general statements of the intended outcomes. Eisner (1967) points out, for example, that ". . . curriculum theorists have tended to neglect . . . the difference between defining an objective and establishing a direction. In defining an objective, the particular type of desired student behavior is described in advance. . . To establish a direction for inquiry, dialogue, or discussion is to identify a theme and to examine it as it unfolds through the process of inquiry." Krathwohl (1965) seems to be saying about the same thing when he suggests that objectives should be specified at several levels of generality.

Equally important, I think, is for the professional evaluator to recognize that the formulation of program objectives is primarily his responsibility. It is true that he must seek the cooperation of the curriculum specialists as he attempts to define these goals. Certainly the set of goals which are finally formulated must be acceptable to both the subject-matter specialists and the evaluators as well. At any rate, I think it is poor practice for the evaluator to place all responsibility for definition and formulation of goals on the innovator. A point of view similar to this has been expressed by Scriven (1967).

It is to be hoped also that examination of the goals or intended outcomes of a program will not be confined to student behavior. Skager (1967), for example, points out that ". . .there are many things going on in the school that are highly significant that can be assigned values, but which are only tenuously reflected, if at all, in the learnings of students." A program that did a remarkable job of achieving its intended student behavior goals would be of relatively little value if it turned out to be so distasteful to teachers that they resigned or if the program developed situations which damaged seriously the morale of the entire student body.

Increasingly, there are a number of curriculum innovations in which the content of the program is considered far less important than the processes and attitudes of inquiry that are being developed. A good

example of this type of program is described by Bruner (1966). Tyler (1964) points out that when objectives are being set for a program of this type, it is extremely important to have clearly in mind the nature of the learning process and the method of instruction that is being used. It would be a fatal mistake on the part of the evaluator to concentrate on goals stated in terms of the program content.

Even in a program which is content oriented, the evaluator should try to determine those intended outcomes or objectives which are broader than the immediate program content. Very important learnings are likely to be overlooked if too narrow a point of view is taken in describing the course objectives. In fact, in some instances, a narrow conception of objectives will fail to reveal that the program is succeeding at the expense of other objectives which are of even greater importance to society.

Let's go back to Cronbach's definition: Evaluation is " . . . the collection and use of information to make decisions about an educational program." In order to provide some background for a discussion of objectives, I tried to indicate the kinds of decisions toward which evaluation is directed. You will recall that these were divided into three general categories: feedback for program development, judgments about existing programs, and instructional research. The basic activity of the evaluation process is that of collecting useful data--data relating to antecedents (conditions prior to treatment), transactions (what goes on during the program), and outcomes (the results of the treatment). Within this framework we must also collect information about the objectives or intentions of the program developer. In many instances, a general rationale for the program is sufficient to get things off the ground, but the evaluator can serve an extremely important function by helping the innovator keep track of the extent to which the activities he proposes to include in his program move him closer to the objectives he wishes to achieve.

Specifically how all these observations concerning antecedents, transactions, and outcomes as well as the information about objectives are used, processed, and interpreted to arrive at the three types of decisions will be presented by some of our other speakers.

REFERENCES

Atkin, J. M. Using behaviorally-stated objectives for designing the curriculum: a cautionary note. The Science Teacher, May, 1968, 27-30.

Bruner, J. S. Toward a theory of instruction. Cambridge, Mass.: The Belknap Press, 1966.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record. 1962, 64, 672-683.

Eisner, E. W. Educational objectives: help or hindrance, The School Review. 1967, 75, 250-260.

Eisner, E. W. A response to my critics, The School Review. 1967, 75, 277-282.

Glaser, R. & Reynolds, J. H. Instructional objectives and programmed instruction: a case study. In C. M. Lindvall (Ed.) Defining Educational Objectives. Pittsburgh: University of Pittsburgh Press, 1964, pp. 46-76.

Hastings, J. T. Curriculum evaluation: the why of the outcomes. Journal of Educational Measurement, 1966, 3, 27-32.

Krathwohl, D. R. Stating objectives appropriately for program, for curriculum, and for instructional materials development. The Journal of Teacher Education, 1965, 16, 83-92.

McGuire, Christine H. A proposed model for the evaluation of teaching. In Ed. 2 The Evaluation of Teaching. Washington, D. C.: Pi Lambda Theta, 1967. Pp. 94-97.

Scriven, M. The methodology of evaluation. AERA Monograph Series on Curriculum Evaluation, 1967, 1, 72-80.

Skager, R. W. Are educational researchers really propared to evaluate educational programs? In Proceedings of the Educational Testing Service Western Regional Conference on Testing Problems. Berkeley, Calif.: Educational Testing Service, 1967, P. 40.

Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 527-532.

Stake, R. E. A research rationale for EPIE. The EPIE Forum, 1967, 1, 7-15.

Tyler, R. W. Some persistent questions on the defining of objectives. In C. M. Lindvall (Ed.), Defining Educational Objectives. Pittsburgh: University of Pittsburgh Press, 1964, Pp. 81-83.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: nonreactive research in the social sciences. Chicago: Rand McNally & Company; 1966.

Wilhelms, F. T. (Ed.) Evaluation as feedback and guide. Washington, D. C.:

RESEARCH DESIGN AND THE INTERPRETATION OF RESULTS

George Temp
Research Psychologist
Educational Testing Service

I would like to sketch for you in the next few minutes some of the underlying principles and problems of research design, not only as they pertain to the evaluation of educational innovations but in the larger context of science: in the context of securing dependable knowledge about any aspects of a confusing and uncooperating Natural World.

To move quickly to the central concern of evaluation, I would suggest the oversimplification that research design is concerned with only one thing:

> how to gather information that will allow one to answer specific questions important to some person, project, school system, decision-making body, or theory.

The problems of research design, and thereby interpretation, arise from two well-established facts:

(1) Individuals vary one from another in a number of significant dimensions.

(2) All measurement procedures in all sciences are subject to some unknown amount of random error that often can exceed any expected effect associated with an innovative procedure.

Let me illustrate briefly the nature of these problems. Imagine for the moment that these two facts were not true. That is to say, imagine that individual members of a definable group did not vary and that measurement devices were perfect measures. Do you see how simple things would be? You could take any one individual, apply your innovative procedure or treatment, and know precisely what the outcome would be if you applied the same procedure to all other members of the group.

For instance, if someone came up with a new idea for teaching reading, we could take any five- or six-year old and apply the procedure, see how well the new method worked, and compare the results directly with those of the old method which, in this imaginary world, would also be one value for all individuals on any specific measurement device.

Some sciences approximate this imaginary state of affairs. In particular, the physical sciences have been able to eliminate a great many but not all error-of-measurement problems by the use of highly refined measurement procedures. Thereby, it is possible to detect very small changes associated with treatments of individual physical structures. In addition, a great deal of time and effort is spent in obtaining analytically pure chemicals or in isolating "pure" cases in the investigation of physical phenomena.

All of this is done in order to control individual variation and reduce measurement error or, in other words, to deal with the problems of research design.

In the science of human behavior--and especially in the area of education--the basic problems of research are the same, but the required solutions are different. The solutions are different because it is impossible, as well as undesirable, to eliminate individual variation in response by some "refinery" process or by isolation of "pure" cases. And the solutions are different because the rights and preferences of humans must be considered in developing procedures attempting to reduce measurement error. It is one thing to split an atom to examine the nucleus and quite another to subject a person to a stress interview and thousands of questions to determine his personality style, although such a procedure might be quite necessary from a measurement point of view.

And, although it may not require repeating, any attempt to isolate "refined" or "pure" cases is doomed to failure. If almost no two individuals have the same fingerprints, then what are the chances of isolating cases that are "essentially" identical and therefore any one (or any small number) can tell us all about the others?

The solutions are different, therefore, not because they are not the right solutions, but because the nature of things frequently blocks us from using direct solutions. Parenthetically, things are not

really as simple or direct in the other sciences as they may seem in this sketch: there, too, researchers must resort to solutions very much like those I am about to present. We are faced, then, with two problems in research design: (1) individual variation, and (2) measurement error. And we are blocked from solutions found effective in other highly successful scientific activity. How, then, are we to proceed?

Unfortunately, now that you have indulged me this far, I shall repay your attention with a dash of cold water. Specifically, although I shall now describe some proposed solutions to these research design problems, I must be quite frank and report that few of the solutions have either the simplicity, robustness, or elegance that stirs one to action and admiration. Also, few of the suggestions have thus far produced even a small hill of information about effective educational treatments, much less the needed mountain. In my own work I am now much more modest; I endorse the approaches described below as capable of producing information of value in optimizing decisions but I cannot say they are infallible, or the best we will develop given more time and experience.

PROCEDURES IN RESEARCH DESIGN AIMED PRINCIPALLY AT OVERCOMING PROBLEMS OF INDIVIDUAL VARIATION

Of the procedures designed to overcome the problem of variations in individuals, four seem especially worth noting: randomization, blocking, covariance, and factor combinations.

Randomization

The procedure of randomization with its implied corollary, large numbers of individuals, is the first solution that comes to mind in designing research studies. Whenever individual variation might lead to wrong conclusions, the procedure of randomization serves to free the investigator to proceed. How does randomization achieve this magic?

Without presenting the mathematical basis for the reliance scientists place upon randomization, I believe I can illustrate how such a procedure works and what benefits it has. The handout (Figure 1) entitled, Evaluation Game-Summary of Rules, assigns a number to you. Please look in the upper right-hand corner of your sheet. (Although we will not be able to cover all of the "rules" presented on this sheet, I believe most of the points mentioned during this conference are implied in the summary statement.)

Now, for a brief audience participation game-

> (Various groupings of the audience illustrate that random assignment gives roughly comparable groups of people on several easy-to-see variables, e.g., sex, bald heads, certain color clothes, etc.)

In summary, I believe this audience participation game illustrates how randomization gives us groups that do not vary significantly from one another, in a mathematical as well as a practical sense. Therefore, although the individuals have not been changed, randomization allows us to look upon the mean values of the groups as comparable numbers. Now, if we introduce some innovative treatment to one group while withholding it from the other, logically we may compare the resulting measured effects with confidence that individual variation alone could not account for any observed differences. Such a widely valid research design is represented on your handout (Figure 1). Several statistical procedures may be applied to the purely descriptive statistics you would use to summarize the outcomes of the above experimental design in order to aid you in deciding whether or not the two obtained means are really different.

If you have grasped the underlying benefit of randomization--the effective neutralization of individual variation by a grouping process that averages out such variation over a sufficient number of cases--then its use in the design of a specific research study should be easier. Ready-made designs that use randomization at various points are described in Campbell and Stanley (1963), Cochran and Cox (1957), Cox (1958), Edwards (1962), and Lindquist (1953).

Blocking

Another response to the problem of individual variations has been to attempt to group into sets--within the limitations of time, money, and knowledge--those individuals who are as much alike as possible. These sets, often called blocks in research design books, are then randomly split and the treatment applied and analyzed as above or, if possible, both the new and the old treatments are applied to the entire block and all analyses are made within a set. The benefit of this blocking is dependent upon our knowledge of good ways to block--e.g., by levels of

Table of Random Numbers (0-9)		
2	7	8
9	6	7
6	7	3
3	8	8
8	3	6
3	2	3
8	6	8
3	8	5
5	9	4
8	1	6
7	5	7
8	0	6
1	6	8
6	3	9
8	7	0
8	6	3
9	2	5
2	6	8
5	4	7
8	4	3
3	4	9
0	6	6
2	7	9
5	7	2
5	6	2
6	9	3
7	0	0
6	0	6
7	6	8
1	5	1
3	3	9
1	6	0
5	0	7
7	6	1
9	2	6
4	1	7

EVALUATION GAME - Summary of Rules

- Definition of Objectives
- Acceptance or Development of Measures
- Definition of Comparison
- Inferences and Decisions

. . . . . .

Common Threats to Reasonable Evaluation Conclusions

Comparison Weaknesses

No defined comparison

Comparison not appropriate

Comparison undefinable

Criterion Weaknesses

More than one superlative at a time

Flexible interpretation of criterion

Misplaced confidence in appropriateness of criterion

Inferential Weaknesses

Interaction of components

Time-experience continuum overlooked

Non-generalizable to desired group

Alternative inferences not evaluated

. . . . . .

A Widely Valid Research Design

A randomly constituted data source — Experimental treatment — Post-measures

Another randomly constituted data source — Non-experimental treatment — Post-measures

(Logically, appropriate source of data depends upon possibility of independence of response to treatment)

. . . . . .

Common Statistical Tools

Descriptive: frequency (enumeration); mean; standard deviation.

Inferential: standard scores; t ratios; F tests.

Bayesian: prior probabilities; orderly revision of opinion.

intelligence or prior achievement. The principal advantage of blocking is that it allows us to make more precise comparisons by eliminating some of the individual variation without increasing the number of students.

There are a number of ways to block, and appropriate analyses that may be used with each. The references mentioned earlier develop the alternatives more fully. If you remember that blocking may help to get a more precise comparison with various groupings of individuals, and without necessarily increasing the numbers, then you can examine these references when designing a specific curriculum evaluation.

Covariance

Covariance is a response to our desire to eliminate or control individual variation when gathering information that will allow us to answer specific questions of importance. As such, it is related to blocking. However, covariance may be used instead of blocking, or along with blocking, in order to make possible more precise comparisons. Essentially, covariance is nothing more or less than a statistical method for adjusting outcome scores of individuals by taking into account initial differences of individuals on one or more concomitant observations. Because so many attributes of individuals are correlated or in essence interacting, it becomes possible to mathematically manipulate the scores of individuals to take certain of these relationships into account.

The best illustration of the use of covariance to eliminate individual differences comes, of course, from the handicapping system used in golf. Here, crudely but effectively, a person's past performance is used to adjust his today's score, so that players of widely differing past performances may play an exciting contest. Covariance analysis does this adjustment more precisely, and sometimes with more variables than just past performance on the same task. Any number of related variables may be used to make more precise comparisons possible. The references mentioned earlier also cover the use of covariance in research design and interpretation. A step-by-step application of covariance analysis is available in Dyer and Schrader (1960).

Factor combinations

I would like to introduce briefly a fourth response to the problems of individual variation. This procedure is based on the fact that we frequently find that different individuals respond to different treatments in different ways. At times we would like to examine these interactions to see if our answers to specific questions can or ought to be made specific to treatments and individuals. Often, of course, we are unable to see how such information could be utilized in a particular school system and, therefore, we use simpler and more direct comparisons of means of class-size groups or other practical groupings within the school administrative structure. But even in this case, for your own instruction or for purposes of suggesting revisions in school practices, you may wish to examine, where possible, the obtained effects where certain interacting factors (such as sex and intelligence level under different treatments) cause significant differences in desired outcomes. The use of factorial research designs is the answer because it allows such interactions to be revealed and, in this sense, eliminates the problem of individual differences by investigating a limited set of such significant dimensions. In certain of these designs, an individual may actually serve as his own control. Unfortunately, the study of possible factorial designs is not entered lightly, and you will probably need the advice of a trained research person in order to use such designs. The references mentioned earlier discuss the procedures and analysis of such designs.

We now leave the procedures designed to overcome the first problem of research design and turn to the second: problems arising from the fact that all measurement procedures are subject to various kinds of error.

PROCEDURES IN RESEARCH DESIGN AIMED PRINCIPALLY AT OVERCOMING PROBLEMS OF MEASUREMENT ERROR

I would like to review briefly four decision areas designed to improve measurement procedures. These are: choice of experimental units, number of observations, nature of observations, and stability of observations.

Choice of experimental units

If you look again at your handout (Figure 1), you will notice a parenthetic remark under the description of a valid research design.

This sentence states that an appropriate source of data depends upon the possibility of independence of response to the treatment. This comment was meant to call attention to the need to consider just exactly what is the experimental unit in any research design. Is the unit the individual student, the intact class, a grade level, a school system, or what?

Although inconvenient and often even hard to detect, there are forces operating in many curriculum innovations that make the logically appropriate unit very large indeed. Let me illustrate with a common problem before indicating how I believe the decision affects measurement error. A teacher of a class of 8 or 80 makes a decision to have more class discussions. Logically, regardless of the number of pupils in the class, the unit used in the experiment is the entire class, because the treatment is applied to all, and each student changes the nature of the treatment by his response or lack of response in the discussions. One student, a good talker, may make or break the discussion procedure. Therefore, it seems clear that no individual student but only the class as a whole can respond to this innovation. As a matter of fact, teachers often comment about the class responding well or ill to certain changes she introduces. This implicit awareness acknowledges that the true experimental unit here is the class, which therefore makes only the mean an appropriate measuring number; an individual score is impossible to evaluate as to measurement error and therefore is valueless until additional units are added to overcome the problem of individual variation mentioned earlier.

Perhaps this is a good point to call attention to the somewhat artificial separation of design considerations employed in this presentation and all others, for didactic reasons. Obviously, the design of any study requires an interplay and review of all decisions, since each decision sets a limitation for other decisions.

The advantage of experimental units whose value is determined by a number of individual cases, although these scores are not used in an analysis, is that the mean is a number that will not change much on most measurement procedures. Thereby, the total number of experimental units required to make a powerful test of an innovation will be less than when individual students are the unit. This means if the class is

the unit, 10 class means are much more stable numbers with a shorter range than 10 individual scores. In effect, measurement error has been overcome by the choice of experimental unit.

Number of observations

Mention has already been made of the number of observations as a variable in the design of research studies. Beyond the sense of number of observations used in discussions of sample size, there is the sense in which number of observations refers to the amount of observation per individual. In effect, what is accomplished by increasing the number of observations per individual is what is accomplished when the experimental unit is defined as a class. That is, a number of observations are used to get a more stable measurement of each unit. Therefore, the comparisons of interest are made among values that have had some measurement errors removed by a repeated observation process. Of course, this desirable goal is often not obtained because of costs in time, money, and irritation of the individuals concerned. However, many opportunities arise when repeated measures may be obtained to good effect. In the discussion of the nature of the observations, to which we will now turn, mention will be made of obtaining different but repeated measures and some of the attendant benefits.

Nature of observations

One of the most significant ways of reducing problems of research due to measurement error is by a detailed consideration of the nature of the observations to be made. Obviously, if the measures do not fit the research questions to be answered, unnecessary and difficult-to-eliminate error creeps into the measurement process. Other sessions at this meeting are devoted to details of the criterion problem, both in terms of the usefulness of available instruments and the problems encountered in the construction of new measurement techniques for first use in a local evaluation effort. I can only emphasize here that thoughtful consideration of the total criterion problem will also profit from discussions centering around the question, "What observations would convince others unfriendly to the program that such and such an objective has been accomplished?"

I would also recommend to you the publications by Metfessel and Michael (1967) and Webb, et al (1966) for descriptions of different kinds of observations that are often first considered.

Stability of observations

Fundamental to the reduction of problems of research design caused by measurement error is the elimination of as much as possible of the instability of observation due to various causes, and referred to as unreliability. There are at least two main methods of increasing the reliability of most observation procedures:

(1) increase the total number of valid observations, and

(2) reduce ambiguous observations or ambiguous procedures.

A helpful discussion of reliability will be found in Educational Measurement (Thorndike (1951); revision in preparation). There you find emphasized the need to look carefully at measurement procedures for signs of loosening and slipping that allow unreliability to creep in. In particular, many measurement procedures are so unreliable that they are of no use. Rubber rulers that give different values every time one measures a room are no help in ordering wall-to-wall carpeting. Yet, with even a touch of stability in a measurement procedure, we profit a great deal. This is why the king's hand, and the king's foot, and the king's thumb helped physical measurement get off to a start that has led to the use of the oscillations of atomic structures for certain present-day measurement procedures. By considering alternatives to insure that we are measuring something as reliably as we can at the present time, we overcome a major source of error that leads to the need for complex designs and complex analyses to uncover relationships.

Summary

By way of summarizing the remarks of the past few minutes, I must tell you of some emerging criticisms of the kinds of research design strategies presented above and in the references cited. In particular I would recommend to you the work of Daniel Strufflebeam (1968) and Egon Guba (1965). Briefly, and perhaps unfairly to them, they seem to be saying that present solutions to the two problems of research design I pose (although they do not use my language) have been unsatisfactory. In

place of these solutions they would recommend a continual flow of information collected on site, with a great deal more reliance placed upon judgment and decison-making in the process. Perhaps from my account of how I view the problems, you may see that I also emphasize decision-making based upon evidence.

Thus, I believe the ultimate question of research is still: How do we gather information that will allow us to begin answering questions of interest?

I doubt that we have the final answer as yet. I hope that the next time we meet we will be farther along.

REFERENCES

Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook for research on teaching. Chicago: Rand McNally, 1963.

Cochran, W. G. and Cox, G. M. Experimental designs, New York: Wiley, 1957 (2nd ed.).

Cox, D. R. Planning of experiments, New York: Wiley, 1958.

Dyer, H. S. and Schrader, W. B. Manual for analyzing results of an educational experiment (analysis of covariance). Princeton, N. J.: Educational Testing Service, 1960.

Edwards, A. L. Experimental design in psychological research, New York: Holt, Rinehart & Winston, 1960.

Guba, E. G. Methodological strategies for educational change. Paper presented at the Conference on Strategies for Educational Change, Washington, D. C., November 8-10, 1965.

Lindquist, E. F. Design and analysis of experiments in psychology and education, New York: Houghton Mifflin, 1953.

Metfessel, N. S. and Michael, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs, Educational and Psychological Measurement, 1967, 27, 931-943.

Stufflebeam, D. L. Evaluation as enlightenment for decision-making. Paper presented to the Working Conference on Assessment Theory, Association for Supervision and Curriculum Development, Sarasota, Florida, January 19, 1968.

Thorndike, R. L. Reliability In E. F. Lindquist (ed.), Educational Measurement, Washington, D. C.: American Council on Education, 1951.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive measures: nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

USES AND LIMITATIONS OF TESTS

Miriam M. Bryan

Senior Editor and Associate Director
Cooperative Tests and Services

Some of you may have seen the little squib that appeared in an issue of The Reader's Digest shortly after the Supreme Court decision regarding prayer in public schools, in which a small boy is reported to have said to a classmate, "It may be unconstitutional but I always say a prayer before I take a test." And well he might -- because the course of his school career -- indeed, the course of his whole life -- may be affected by a single test score or set of test scores.

Testing has been variously described as tyranny, as a menace to education, and as a necessary evil. It is necessary, but it need not be tyranny, a menace, or an evil. Testing can be a beneficent tool of education when school people understand what tests can do and what they cannot do, and when they learn how to select tests, construct tests, give tests, and use test results cautiously and wisely.

The ultimate goal of any good testing program is the improvement of instruction, brought about by teaching each student in the way best calculated to develop his abilities to the limits of his potential. To accomplish this goal, a number of different kinds of measures are needed -- aptitude and achievement tests, survey and end-of-course tests, standardized and teacher-made tests -- along with information about personal characteristics and interests that may influence the direction that the instructional program will take. In order to achieve the ultimate goal, the good testing program will serve several intermediate purposes, with test results used for making decisions about placement, grouping, and promotions; for identifying and diagnosing sources of learning difficulties; for assessing the effectiveness of teaching materials and teaching methods; for making predictions about future performance; and for evaluating the total school program.

Most frequently the tests that are used in decision-making are made by teachers. Standardized tests become important when the reference group must be larger than the class group -- when it is essential to know, for example, how the amount of learning that has taken place in a particular classroom or school compares with the amount

of learning that has taken place in similar classes or schools elsewhere, or how an individual's present ability to perform compares with that of large numbers of others who aspire to pursue the same course of study or to engage in the same career.

At the present time, there are standardized tests to fill almost every need for standardized testing. The Sixth Mental Measurements Yearbook lists 1200 of them. Hundreds of these are, in the opinion of the experts who have reviewed them for the Yearbook, quality tests. They are quality tests in spite of limitations noted by experts in their reviews, that may be unique to each test or series of tests. Why, then, with so many tests available and with their limitations so well delineated by the Yearbook reviewers, do we need to talk about limitations here? I suggest that this is because there are limitations other than those inherent in the tests themselves: namely, the false assumptions held by test users about what tests can do.

I should like to present to you several false assumptions that seem to me to stand in the way of good testing. I think these assumptions are responsible to a large degree for the mistakes that are made in the selection of tests, in the construction of tests, and in the use and interpretation of test results, and for the erroneous impressions that are thereby created among students and teachers and the public at large.

A first false assumption is that aptitude and intelligence tests currently available measure some inborn ability that determines for his lifetime an individual's potential for learning. There is little doubt that general ability is linked to some extent to heredity. This inborn ability does exist, but aptitude and intelligence tests, as they have been developed, do not and cannot measure it -- at least not in such a way that we can currently interpret results of such tests with any confidence.

Aptitude and intelligence tests purport to measure the individual's capacity to learn. What they do measure is his ability to perform certain kinds of mental tasks. They measure this performance not at birth but a long time afterwards. The tasks are the kinds that the individual learns to perform as a result of his experiences at home, in school, on the playground, on the job, and elsewhere. The amount of learning gained from these experiences varies greatly from one individual to another. It varies with the regard in which learning is held in the individual's home, the language he hears at home and among his peers, and the quality of the instruction

offered by the schools he attends. It varies as the environments of individuals vary. A score on an aptitude or intelligence test cannot bypass all the experiences that help or hinder an individual's learning. A score on an aptitude or intelligence test cannot be interpreted as reflecting directly the extent of the brainpower with which he was endowed.

While we think of the aptitude or intelligence test as something different from the achievement test, the two kinds of tests differ less than is commonly supposed. Aptitude and intelligence tests are intended to reflect the amount learned from incidental experiences before special training is received; the abilities tested are presumed to be common to individuals regardless of home or school. Achievement tests are intended to reflect the amount learned in school. In both kinds of tests, the abilities tested are products of the individual's inherited potential for learning and his opportunities for learning. The main difference between them is that the tasks the individual is required to perform on an aptitude or intelligence test are learned over a relatively longer period of time than those he is asked to perform on an achievement test. They are able to predict an individual's performance not so much because they measure his inherited ability to learn, as because they show what he has learned in the past that will help him learn in the future.

A second false assumption is that a test score is perfectly reliable -- that a score made by an individual on a test today is the same as the one he will make tomorrow or next week on the same or a similar test. It sometimes does not occur to those without training in tests and measurement that a test is only a sample of an individual's performance, and that it is a sample that can never give any more than an estimate -- sometimes a very rough estimate -- of how much he knows about what is being measured.

In school testing situations, the test score may be affected by a multitude of things: by the atmosphere of the room in which the test is given, by the personality of the teacher who administers the test, by the school or the family social calendar for the week, and even, perhaps, by what the student has for breakfast on the day the test is given. The test score is affected most seriously, however, by the questions that are asked. Mary in grade 4 may be able to spell "though" but she may not be able to spell "through;" John in grade 5 may know that the capital of Virginia is Richmond, but he may not know that the capital of Washington is Olympia. All tests have a standard error of measurement that makes the score obtained

very seldom the true score, i.e., the average score that the individual would make were he given many tests in a given area rather than only one test. All the things that I have mentioned contribute to that error, and especially the questions asked.

I can illustrate the standard error most easily with reference to IQ scores. The Revised Stanford-Binet Intelligence Scale, which most psychologists would classify as one of the most reliable tests ever developed, has a standard error of 5 IQ points. What should that mean to the individual interpreting the test scores? Well, with an obtained IQ score of 100, there are 2 chances in 3 that the true IQ score falls somewhere between 95 and 105, but there is one chance in 3 that it does not. There are 95 chances in 100 that it falls somewhere between 90 and 110, but there are 5 chances in 100 that it does not. But we can be almost certain that it falls somewhere between 85 and 115. So what do we have here? -- a possible range from "dull normal" to "above average." This is the degree of accuracy with which we can describe a score on one of the most reliable tests we have!

I say again that what I have said about the standard error of measurement applies to all tests -- aptitude or intelligence tests and teacher-made or standardized achievement tests. Did you know, for example, that on a 100-question classroom test the standard error is likely to be about 5 -- which gives the same spread of possible true scores that I described for the Stanford-Binet? If a student answers 80 questions correctly, we can only say with almost certain assurance that his true score falls somewhere between 65 and 95.

Knowing something about the standard error should make school people hesitate to label a student with a particular IQ on the basis of a score on an intelligence test. It should make school people hesitate to report that students achieving particular scores on standardized tests are performing at particular grade levels. It should make school people hesitate to say that one grade on a teacher-made test is passing and that another is failing. Test scores are not reliable enough to permit these conclusions to be drawn.

A third false assumption is that standardized achievement tests should measure everything in the subject-matter areas with which they are concerned. Tests cannot measure an entire universe of subject matter; they can measure only samples of it. Furthermore, since standardized

achievement tests must provide measures of the amount of learning that has been accomplished in a wide variety of learning situations, they cannot provide thorough and complete measures of what is covered in specific textbooks or specific courses of study. Critical examination of the content of any standardized achievement test will reveal that even though the test may be the one best suited for a particular purpose, it still will not sample all aspects of the subject matter that have been emphasized, and it is very likely to be concerned with some aspects of the subject matter that have received no emphasis at all. Interpretation of scores on standardized achievement tests, without knowledge of just how valid the test is for the particular testing situation, may result in completely erroneous conclusions about the significance of the student's score.

Several years ago I was employed by a textbook publisher who brought out a new series of arithmetic books for the elementary grades, in which certain arithmetic concepts and operations were introduced in different grades than they had been in the older series. Before the series was more than two years old, we began receiving letters from school people reporting that their students were <u>not</u> performing up to grade on standardized arithmetic tests and asking why our series was not producing better results. When we inquired what tests they were administering, we found that the tests had been written and standardized fifteen years before the textbooks were published. When we examined the tests with them question by question, and indicated for each grade level the questions that students using our texts should be able to answer and those questions they could not possibly be expected to answer, the school people found that their students were performing quite well. The administrators had selected the tests without examining them with the course of study in mind.

There are many, many mistakes made in the selection of standardized tests for school use. A student in one of my college classes reported a problem that had become serious in her school, and for which nobody in the school had found a satisfactory answer. In that school, reading tests were administered in the fall and in the spring. Each year the grade-equivalent scores achieved in the fall were higher than those achieved the following spring. As far as the teachers and the principal could see, the children were falling back in reading through the school year and making remarkable recovery during the summer. What was happening in the classrooms

in that school during the year? Actually, nothing bad. The principal had found some old tests in the basement and had decided to administer one reading test in the fall and a different one in the spring -- two different reading tests standardized on different populations. Scores on two different achievement tests, written by different authors and standardized on different school populations, just cannot be compared.

A fourth false assumption is that a student's scores on a battery of achievement tests give all the information that one needs to make decisions about what and how much a student has accomplished as a result of the learning experiences he has had in the past, and what and how much he will be able to accomplish as a result of the opportunities for learning that he will have in the future. No test battery currently published can do this. Tests can show the student's strengths and weaknesses in the various subject-matter areas tested, and they can show how he stands in these areas when compared with his peers, with individuals in the reference groups, or with reference groups as a whole. But there are many important outcomes of learning that cannot be measured by any test battery so far devised -- outcomes that are deeper in origin and greater in penetration than any revealed by test scores -- outcomes that can only be evaluated by the human beings who are able to observe the student closely and over a long period of time: his family, his teachers, his peers.

Closely related to this false assumption is a fifth one: that a profile of scores on a battery of achievement tests presents a considerable amount of reliable information about the strengths and weaknesses of the student in several different subject-matter areas. It does not necessarily do this. The differences that the profile shows, even though they appear to be large, may not be reliable differences for a variety of reasons: the scores plotted may not be true scores; the score scales for the several tests represented on the profile may not be comparable; the tests represented on the profile may have been normed on different populations; and the several scores shown on the profile may not be independent measures but rather highly correlated measures. As long as I have been in testing and long before that, I am sure, test specialists have been worrying about the problem of the profile because it offers possibilities for all sorts of errors in score interpretation. The profile is still in wide use, however, because it presents a picture that appears to be easy to read and convenient to explain.

School people need to be especially careful in their interpretation of profiles based on performance on achievement test batteries. I do not know how many school people have said to me, "We have made a fine record for ourselves in reading. Our sixth-grade class is reading at the 8.2 grade level. We haven't done so well in arithmetic: our sixth-grade class is achieving at only the sixth-grade level." They do not stop to think that progress in reading is -- or at least should be -- continuous, while progress in arithmetic depends almost entirely on what is taught in the schoolroom. The child who learns to read the words "mad" and "hat" can usually very quickly read "had" and "mat," but the child who has learned to multiply with a one-digit multiplier cannot multiply with a two-digit multiplier until he has been taught how to do it. Above the very lowest grade levels, I would <u>always</u> expect grade scores in reading to be higher than those in arithmetic and other content subjects.

A sixth false assumption is that grade equivalents on standardized achievement tests give an accurate and easily interpretable picture of the level of a student's performance. Grade equivalents imply that boys and girls progress at an even pace through the school year, and that they do no forgetting over the summer. We know that neither of these implications is true. Children progress at sharply irregular rates through the school year, and they do so much forgetting during the summer that in some subject areas, like arithmetic, it is late October or early November before they are doing work of the kind they were doing the preceding May. But have you ever seen tables of grade equivalents that reflected this?

Test publishers have perpetuated grade equivalents not because they think they are particularly accurate, but because test users think they are easily understood by teachers and easy to explain to parents. Perhaps the test publishers have not made it clear enough to test users that except for grade equivalents which coincide with the time of year when the norming was done, all the rest of the scores are estimated. No publisher I know tests children every month of the year to obtain their grade equivalents. They find the difference in the average scores for two adjoining grades at whatever time of year the testing is done, and then divide that difference into ten equal parts to which they assign grade equivalents.

And what does a grade equivalent mean, anyway? If Jimmy in grade 3 achieves a grade equivalent of 5.5 in arithmetic, should he be transferred to a fifth-grade class? No, because he has not had the fourth-grade work

and is totally unprepared to carry on at the fifth-grade level. The 5.5 grade equivalent simply means that Jimmy is doing well in grade 3. A much more meaningful description of his score would be that it has a percentile rank of 94. That is, his score is higher than the scores of 94 percent of the third graders in the norms group.

A seventh false assumption is that a norm is a standard -- that it represents just what a student or a group of students should be achieving at a particular time. A norm does not tell us anything about what students should or should not know. It simply describes the performance of the group of students who took the test in the standardization program. If the test has been normed on a group of low achievers, it is easy for a student of average ability to get a score that is above the norms group average; if it has been normed on a group of high achievers, however, such a student is not likely to get a score that is above the norms group average. If I were a teacher, I would want to think very carefully about the general ability of my students compared with the ability of the norms group before I decided whether or not I was happy with my test results. If I were teaching a very bright class, I would be unhappy if very many of my students fell below the average for an average norms group. If I were teaching a group of low achievers, I might be very happy if only a few of them came almost up to the norm for an average norms group. Norms have to be interpreted in terms of the general level of the group being tested and the ability of the students comprising the norms group.

More and more we are urging school people to use national norms simply as reference points when they want to check the level of achievement or performance of particular class groups against that of other school groups. For a more useful evaluation of their test results they should compare the performance of their students with that of other students in the same school and in the same city. It is relatively easy to develop local norms; you can do it with a very modest background in statistics.

School people need to be very much concerned about the pitfalls of score interpretation. I have mentioned the inadequacy of grade equivalents. I have also suggested that national norms are frequently less suitable for score interpretation than are locally constructed norms. I would now like to caution you about three other hazards of score interpretation.

First, the norms accompanying most standardized tests are based on the achievement of individual students; they are not based on the average achievement of groups. They should be used, then, in interpreting the scores of individual students; they should not be used in interpreting class or school averages. A high group average for a class will look lower than it should if it is interpreted using individual score norms; a low group average will look higher than it should; only a group average close to the mean scores for individuals will look about right. Very few tests have in the past been accompanied by both individual and school norms; it would be helpful if more tests were accompanied by both kinds of norms.

A second caution involves the interpretation of scores on objective tests that use a correction-for-guessing scoring formula. Sometimes we do not stop to think that on a 50-item true-false test a student may get a score of 25 by chance alone and that on a four-choice multiple-choice test of 100 items he may also get a score of 25 by chance alone. The advantages and disadvantages of correcting for guessing will not be discussed here -- too many test specialists have been arguing these for too many years, without reaching any kind of agreement. The caution must be expressed, however, that when objective test scores that have not be corrected for guessing are examined, the score that can be achieved by chance alone should be considered carefully in the interpretation of the scores. If using a correction-for-guessing formula has any merit, then uncorrected scores at and below the chance level should be considered zero scores.

A third caution to keep in mind in the interpretation of test scores is that in any testing situation students are working under pressure, if they think the test score is of importance. The scores they make, therefore, are more likely to be indicative of their frustration levels than of their instructional levels. If I were grouping students for instructional purposes on the basis of scores on a standardized reading test, for example, I would want to think seriously about the suggestion made by Emmett C. Betts, a noted reading specialist, to the effect that students should be instructed at a grade level below that indicated by their test scores, and should be offered supplementary experiences at a level lower than that.

I indicated earlier that in order to plan each student's learning program to develop his abilities to the limits of his potential, information about his personal characteristics and interests is needed along with scores

on different kinds of tests. I did not mention personality tests and interest inventories as such because, in my opinion, most measures of this kind so far published have had so many inherent limitations that their usefulness has been questionable. I have seen one personality test after another announced with much fanfare one year and then quietly withdrawn from distribution a few years later. Many interest inventories have suffered the same fate. Only the hardiest of these instruments have managed to survive for any length of time -- and even these have had their shortcomings. An eighth false assumption would be, then, that personality tests and interest inventories, as they are now conceived and constructed, can offer the kinds of evidence on which decisions concerning course of study or career may be made with some degree of confidence.

What is there about personality tests that has rendered them so generally unsatisfactory to date? First of all, they have been controversial measures because psychologists primarily concerned with the measurement of personality have not themselves been able to agree upon a universally acceptable definition of personality. If personality cannot be defined to the satisfaction of this group, can it be measured? Second, the statements or questions on most personality tests tend more often than not to be concerned with insignificant and maybe very random behaviors rather than with serious and deep-rooted ones. It has been exceedingly hard for psychologists to design questions for paper-and-pencil tests that really get under the skin. Also, the wide diversity of interpretations to the same responses on projective tests gives cause for concern. Third, there is little evidence that the personality traits that can be measured with any degree of confidence are necessarily permanent traits for the individual being measured. Individuals frequently change as they mature. A very maladjusted first grader may become a quite well-adjusted college senior. Finally, the responses to questions in personality tests can be faked. I have proved this to my own satisfaction by taking the same test several times, and pretending I was a different person each time. I came up with quite different personality profiles every time!

I would suggest that interest inventories suffer from many of the same limitations that personality tests do. While psychologists may have come closer to agreement on the nature of interests than they have on the nature of personality, interest inventories continue to be largely concerned with rather trivial likes and dislikes; the interests of

individuals, especially of individuals of school age, change even more rapidly than personality characteristics; and interest inventory scores can be faked. I have done some experimenting with faking them, too.

In spite of my less than enthusiastic comments about personality tests and interest inventories, I must admit that a great number of students in high school and college have undoubtedly been helped toward decisions about their careers as a result of doing some serious thinking about themselves and their interests. His responses to the questions or statements on a personality test or interest inventory may not tell the student exactly what he should or should not do, but the experience of having to examine himself may give him some perspective about himself and his interests that he has not had before. Whether or not he will be motivated to do something about this is something of an "unknown" even to the individual himself at the present time. Unfortunately, a measure has not yet been devised that will permit a reliable prediction of the extent to which the individual's perspective of himself and his interests will motivate him to develop his abilities to the limits of his potential.

Since I stated earlier that the tests that are most frequently used in decision making are the teacher's own tests, I should like to direct my last few comments specifically to these. There are two assumptions regarding teacher-made tests that seem to me to stand in the way of good testing. These will be my ninth and tenth false assumptions.

A ninth false assumption is that a score on any classroom test can be used with some degree of confidence in reaching a decision regarding the level of a student's performance in the subject-matter area being tested. It is my personal opinion that more sins in evaluating and grading are committed by teachers as a result of over-confidence in scores on poorly planned and hastily contrived classroom tests than as a result of any other contributing factor. A good deal of my professional time is spent going from school to school or from school system to school system to work with teachers on the improvement of classroom tests. I see some very good tests and I see some very awful tests. How the students live through them, and how the teachers can interpret results on them with any kind of confidence is beyond me! I have a collection of test items that are fantastic. Let me give you just a few examples.

Here is a question from an eighth-grade history test that a junior high school teacher thought was measuring critical thinking about life in America:

> Name the presidents of the United States in the order in which they served.

Another question on the same test read as follows:

> Name every member of the President's Cabinet and tell what his responsibilities are.

How many of you could answer these two questions? If you could, would you say you arrived at your answer by thinking critically about life in America?

Here is an essay question on a final examination in American history that twelfth graders were asked to answer in 20 minutes:

> Describe the development of transportation and communication from the period of prehistoric man to the present time.

The teacher who wrote that question had certainly not tried to answer it himself in the amount of time allotted.

How can any answer to this next essay question from a tenth-grade literature test be given anything but full credit?

> Tell everything you know about Charles Dickens.

If a student replies "Nothing," has he not told all he knows!

Here is a completion question from a high school biology test:

> We can taste_____, _____, _____, and _____.

And here is another completion question from a ninth-grade English test:

> ___1___ sent ___2___ to ___3___ to talk with ___4___. Meanwhile four ___5___ were at ___6___ choosing among ___7___ ___8___, which were made of ___9___, ___10___, and ___11___.

Questions like these serve as battles of wit between teacher and student.

I will now cite from a source other than my own collection the best example of a poor teacher-constructed true-false question that I have ever seen:*

> Water boils at 212°.

---

*Tinkelman, Sherman N. Improving the Classroom Test: A Manual of Test Construction Procedures for the Classroom Teacher. Albany, N. Y.: New York State Education Department, 1957.

Superficially that looks true -- and with a poor background in science I would mark it so -- but the good student would not be so sure that it is true, because no information is given about whether the Centigrade or Fahrenheit scale is indicated, what the atmospheric pressure is, and whether the water is pure or contains salts.

This is a multiple-choice question from a ninth-grade civics test that has no wrong answer:

The population of the United States is more than

a. 130 million
b. 150 million
c. 170 million
d. 190 million

There are four right answers. Since the population of the United States is more than 190 million, it is also more than 130 million, 150 million, and 170 million.

The next question represents a kind of error frequently made by teachers when they wish to make certain that most of their students will answer most of their questions. This is from a sixth-grade social studies test:

The Spanish conquistador who conquered Peru was

a. Henry Hudson
b. Jacques Cartier
c. Leif Ericsson
d. Francisco Pizarro

There is only one Spaniard among the men listed -- and only one conqueror as well.

Although matching exercises appear the easiest to construct, they actually present great difficulties to uninitiated test writers who tend to mix up men and mountains, singulars and plurals, different parts of speech, and such, to the point where the student who can read can make the desired choice whether or not he knows anything about what is being tested. Look at this matching exercise -- a favorite of my horrible examples -- from a test on New England poets of the nineteenth century constructed by a high school English teacher:

Column A		Column B	
1.	First line of a poem by James Russell Lowell	a.	Nokomis
2.	Character in Henry Wadsworth Longfellow's "Hiawatha"	b.	Concord Bridge
3.	Author of "Snowbound"	c.	"The snow had begun in the gloaming"
4.	Poem by William Cullen Bryant	d.	John Greenleaf Whittier
5.	Historic landmark referred to in poem by Ralph Waldo Emerson	e.	"Thanatopsis"

And then, after giving tests with questions of this quality, some teachers presume to mark them on a straight percentage basis -- falsely making the tenth and last false assumption that knowledge in the particular subject-matter area can be tested and reported on a 0 to 100 scale.

What does a percent grade on a teacher-made test mean? Does a grade of 80 percent on a test on World War II, for example, mean that the eleventh-grade student who was given that grade knows 80 percent of all there is to know about World War II? Does it mean that he knows 80 percent as much as most eleventh graders know -- or should know -- about World War II? No. It simply means that he knows 80 percent of the answers to the questions about World War II that one history teacher thought it was important to ask his eleventh graders on a particular day, and nothing more than that.

I could spend a whole day -- and more -- talking about the evils of the percentage marking system employed by most teachers. I used to be a teacher and I used to employ it, so I know from first-hand experience what I am talking about. Knowing what I do today about the fallibility of percentage grades, I would never decide what any score was worth, before I examined all the scores on tests that I had constructed and administered. I would expect that none of my tests would be of exactly the difficulty that I intended them to be, that each of them would be somewhat easier or somewhat harder than I had hoped. But perhaps most of you are ahead of me -- perhaps you have abandoned percentage marking long ago.

Decisions must be made about the student's learning potential, and should be made in terms of his demonstrated abilities, his present level of achievement, his personal characteristics, and the nature of his

interests. Decisions must also be made about specific courses of study and the general direction of the school curricula, and should be based on the demonstrated accomplishments of many students. These decisions are important not only to the students directly affected by them, but to society in general.

Even though they have some limitations, existing tests can help in making decisions about students and curricula. If tests are properly selected or constructed, administered, and interpreted, these decisions will be more informed and therefore more accurate than if they are made on the basis of purely personal judgments.

In judging the ability of others subjectively, each of us is influenced by his own experiences. Good tests provide a way of going beyond these experiences -- far enough beyond, we may hope, so that no small boy, or big boy, will feel the need to say a prayer before he takes a test.

DEVELOPMENT OF MEASUREMENT INSTRUMENTS: PROCEDURES AND PROBLEMS

Thomas F. Donlon
Director, Test Development Division
Educational Testing Service

I think that given my title, Director of Test Development, and given my employer, Educational Testing Service, you could now logically expect to be at that point in the program where you are presented with a brief account of the procedures you should use in developing instruments for the assessment of the various elusive educational outcomes that your teachers persue. As has been pointed out, standardized testing instruments are available. Frequently, however, these are oriented towards knowledge outcomes. From what you have said this morning, I gather that many of your curricula are oriented toward noncognitive or nonknowledge outcomes. Standardized instruments will therefore probably have at best a limited value in evaluating the effectiveness of your curricula.

Unfortunately, I cannot describe procedures for constructing instruments of the type that will measure some of the noncognitive or nonknowledge outcomes you have mentioned as important objectives of your curricula. As far as I know, such procedures do not exist. When education was expected to produce knowledges, measurement could provide substantial assistance in the assessment of outcomes, but as educational objectives move away from knowledges to attitudes, beliefs, propensities, and so forth, measurement experts begin to stammer a little.

Your needs are real, however, and therefore I would like to talk about three things measurement people have learned about the area in which I say I don't feel confident -- the development of instruments for the assessment of nonknowledge outcomes, attitudes, and so forth.

First, we have learned that behavior indices are clearly more superior to self-report, to teacher ratings, and to peer ratings. The latter three are useful in certain contexts but when the chips are down, behavior is the characteristic to observe in reaching conclusions about the outcomes of education. This may seem so broad a generalization as to be trivial and useless, but I don't think so and neither do most measurement people.

Second, we've come to understand that the results produced by experimentally induced behavior often cannot be generalized to spontaneous behavior, which is frequently the characteristic we wish to observe. For example, if you assemble a hundred students, put them in a library, make notes as to what books they pick from the shelves, and so on, the behavior they are exhibiting may be valueless in predicting what they do in the library when nobody tells them to go there.

The third point is that the endurance of educational outcomes, other than knowledges, is perhaps not as predictable as the endurance of knowledges. That is, an outcome which is an attitude, a propensity, or a belief, will not endure in the same way that a knowledge will: a knowledge will endure or decay over time according to the laws of memory; an attitude or belief is something that may or may not stay with a person, depending on the effects of complex interactions between himself and his environment. This point was touched on by a comment I heard earlier about the extent to which you get marvelous results in the immediate aftermath of some educational treatment, but then when you go back and examine people shortly thereafter, you find that under the influence of a continuing environment, they fail to sustain what you hypothesized for them.

I think these three ideas are about all that we can offer at this time. While I don't think they are the answer, I also don't think they are things to be overlooked.

In the cognitive area, what I propose to do is to present to you three generalizations which currently seem to be accepted by measurement people, and which are clearly relevant to the procedures for assessing knowledge. After these generalizations, I propose to present and discuss some ways of looking at curricula in detail through item analysis. After that, I would like to return to some additional generalizations.

My three generalizations are essentially the same as those Don Melville earlier attributed to Cronbach: we both went to the same well. The article I read was "Evaluation for Course Improvement" by Lee J. Cronbach, in New Curricula, edited by Robert W. Heath and published in 1964 by Harper and Row.

The first generalization is that in studying test results, major attention should be given to individual items or clusters of items, rather than to total scores. I hope to demonstrate later the value of looking at items rather than scores.

The second generalization is that one should maximize the number of different items one uses in assessing a curriculum. If you have 500 students and you give each of the 500 students 50 items, you than have information on each item for 500 students. Statistical sampling studies indicate, however, that you really don't need to know what 500 students do on an item in order to have a pretty good idea what all of the students will do. If you give each item to 100 or 150 students and apply the appropriate statistical techniques, you can make the inference that these randomly selected 100 or 150 are behaving like the total population. Using this technique, you can increase the number of items on which you get data. For example, rather than demanding that all 500 students take all items, you could give each item to only 100 students and thereby get information on five times as many items. This, then, is a powerful way to broaden the sampling of the curricular outcomes you are attempting to assess with your instrument.

As far as content sampling is concerned, the results of this procedure are as validly generalizable to the total content domain as people sampling is to all the people in the population. So, the second premise is that one should maximize the number of items, keeping in mind that each item should be given to a sufficient number of people to provide a reliable estimate of its properties.

Cronbach has also pointed out that one need not limit oneself to items which are directly tied to the curriculum and its objectives. One should try to measure outcomes that are only hoped for, or even dreaded. That is, you can present students with a question or a task for which the curriculum did not specifically attempt to train them. A good example might be in new math, in which only glancing attention may be paid to square root extraction. Under this rubric it would not be inappropriate to present a square root problem to the students -- and don't be shocked if some of them have learned it at home, or if some of them remember the teacher's one-minute or fifteen-minute presentation on this problem.

The third thing that Cronbach pointed out in the article I read was that in this context you can give items which are achieved by 100 percent of the group. The results would be meaningful to somebody in course evaluation work, but not to somebody making an individual comparison test, à la standardized test. The point is that in the absence of the need for an item to contribute to a score, the scope covered by the items can be broadened because, in a sense, the individual items can have poorer technical properties but still provide helpful information. For example, the percent passing a certain item may be zero, but this would still be meaningful to you as a curriculum building or evaluator: it may support your hypothesis that nobody in the group understands the problem. Also, you can use items that are poor for individual measurement purposes because everyone passes them, but which are comforting for course evaluation purposes if the behavior has indeed been taught to the students.

These, then, are the three generalizations: 1) we should examine items individually or in clusters; 2) it is not necessary to have every student take every item; and 3) the appropriateness of an item for curriculum evaluation is not the same as the appropriateness of an item for individual evaluation.

Now, if you will look at Figure 1, on the following page, you will find an item which asks the question, "Which of the following represents Cartesian coordinates?" In true multiple-choice fashion, there are four possible answers from which the student is instructed to choose the one he thinks is correct. This question was put to 1,622 students prior to a course, as a pretest, and at the end of the course, as a posttest. The arrays of numbers below the question provide information on how the students responded to this question.

The data below the question and to the right break down the students' responses into three classes: O for omit, R for right, and W for wrong. For example, across the top you find the numbers 8, 12, and 21, with a Total of 41. This tells you that on the posttest, 41 students omitted the item, and that on the pretest, of these same 41 students, 8 had omitted the item, 12 got it right, and 21 got it wrong.

98. Which of the following represent Cartesian type coordinates?

(A)

(B)

(C)

(D)

98.		Pretest					
		0	A	B	C	D	Total
Posttest	0	8	12	7	6	8	41
	A	39	443	165	186	252	1085
	B	4	19	25	15	20	83
	C	18	76	47	70	83	294
	D	7	23	12	19	58	119
Total		76	573	256	296	421	1622

		Pretest		
		0	R	W
Posttest	0	8	12	21
	R	39	443	603
	W	29	118	349
			573 or 35%	

1085 or 67%

Figure 1.

Thus, running down the side vertically on the left are the posttest results in terms of omits (O), rights (R), and wrongs (W). That is, the 1,622 student responses were broken down into nine subgroups, in terms of whether they omitted the item both times (8, in the upper left-hand category), or whether they got it right both times (443, in the middle), or whether they got it wrong both times (349, in the lower right). The grid also shows how many got it right the first time and wrong the second time (118), and wrong the first time but right the second time (603).

Some of these categories are perhaps more meaningful than others for curriculum evaluation. I think the most meaningful one is the 603 students who got it wrong the first time but got it right the second time. They are probably the students the teacher would point to as the ones who have benefited from her instruction, and it would indeed be appropriate to hypothesize that this group was changed by instruction.

The students who got it right the first time and right the second time might be said to be "neutral." It would be hard for a teacher or an educational process to take much pride in them: they performed at the end of the educational treatment just as they did at the beginning.

I don't know what to say about the 349 students who got it wrong both times, except that this treatment apparently was not sufficient, or perhaps not appropriate, for their particular educational needs.

Then there is a very obnoxious group of 118 students who got it right before the course and got it wrong after the course. All we can say about them is that they probably guessed and, if nothing else, that it is refreshing to be confronted by them in numerical data of this type. They constitute about 20 percent of all the students who got it right the first time. As we move on to results on other items, we will come to one in this category which is very interesting.

Moving over to the left, still below item 98 in Figure 1, we find a more complete display of the same data, which separates the wrong answers. The underlined A at the top indicates that A was keyed as the correct answer: the "Cartesian coordinates" are the cross at the top, which is alternative A. This time the results are categorized into 25 cells, according to whether the examinee omitted the item or answered A, B, C, or D. Now, if you read across

from 0 on the left, you find 12 students chose the correct answer. They are the same 12 students who are described over in the nine-fold table as getting it right but omitting it on the posttest. The 443 students who got it right both times are the students who answered A both times; they appear at the two intersections of A. Column B contains all the students who chose B when the item was part of the pretest administered prior to the course. There were 256 such students, with 165 responding A at the end of the course.

When I total all the groups under columns B, C, and D, I find 256 in column B, 296 in column C, and 421 in column D. This tells me that a grand total of 973 errors were made on the pretest. It seems interesting to ask the question, "What proportion of success was exhibited within each of these three categories on the posttest?" If we go back to the chart, we find that of the 256 students who responded B in the pretest, 165 got it right on the posttest. Similarly, of the 296 students who chose C on the pretest, 186 chose A on the posttest. Finally, for the 421 students who chose D on the pretest, you find 252 choosing A on the posttest. If you express these in proportions, you get 165/256 = .64; 186/296 = .63; and 252/421 = .60. This may be interpreted to mean that the instruction process had about the same success in reaching students, regardless of whether they had misconceptions B, C, or D before the course. That is, about 60% - 65% of all three groups were brought to the correct response.

Another way of summarizing these data is as follows: On the pretest, 76 students omitted the item; 573 chose A; 256 chose B; 296 chose C; and 421 chose D. When we look at the posttest responses, the most dramatic finding is a substantial increase in the number of students who succeeded on the item: the number who chose the correct response rose from 573 on the pretest to 1085 on the posttest. There was on the other hand a substantial decline in the number who chose B -- from 256 on the pretest to 83 on the posttest. The number who chose D also went down. But notice what happened as far as C is concerned: 296 students chose it on the pretest, and 294 chose it on the posttest: as many people chose it the second time as did the first time. This may be saying that there is something in the course that

is advertently sustaining or promoting the error represented by C. Another way of looking at the data is to observe that option C accounted for only 30% of all errors prior to instruction. Following instruction, it accounted for about 60% or double its relative popularity. Why was this? What in the course precipitated this?

To get some clues as to what might be happening as far as the C response was concerned, I read a little about the course that was being evaluated by these data. I discovered that there is a great deal of emphasis on polar coordinates and Cartesian coordinates. It occurred to me that probably some of the students who entered the course had never even heard of polar coordinates, and so C made little sense to them. In the pretest, therefore, they all went for D. By introducing polar coordinates in the course along with Cartesian coordinates, it is possible that the stage was set for confounding the two things being taught, for enhancing the attractiveness of C as a mislead.

This is one example of the use of detailed analyses of pretest and posttest item data to draw conclusions about the success of an educational treatment.

Let's now move on to item 94 which is given in Figure 2 on the following page. If you look at the smaller nine-fold grid on the right, you will see that there were again 1622 students. This time, there were 109 students who moved from the wrong responses at the beginning to the right response at the end, which is a demonstration of learning that must be really heartening if you are a teacher. But if you now examine the number of students who moved from a right response at the beginning to a wrong response at the end, you'll find there were 317 of them. In other words, starting from the pretest success 1380, this group moved through the course to a posttest success of 1174, or a net decline of about 13 percent. This is clearly disastrous in terms of educational goals.

94. Which of the following systems of location does the geographer most frequently employ?

(A) Latitude-longitude
(B) Polar
(C) Cartesian
(D) Celestial

94.		Pretest					
		0	A	B	C	D	Total
Posttest	0	5	24	0	1	1	31
	A	26	1039	34	54	21	1174
	B	4	60	16	8	4	92
	C	7	231	14	25	10	287
	D	2	26	4	4	2	38
Total		44	1380	68	92	38	1622

		Pretest			
		0	R	W	
Posttest	0	5	24	2	
	R	26	1039	109	1174 or 72%
	W	13	317	87	
			1380 or 85%		

Figure 2.

If we go on and examine the more detailed data on the left, as we did for item 98, we find that on the pretest B attracted 68, of whom 34 or or 50% were moved into correctness; C attracted 92, of whom 54 or 59% were moved into correctness; and D attracted 38, of whom 21 or 55% were moved into correctness. When you examine these proportions or percentages, you can speculate on a mild tendency for the students who at the beginning of the course thought that geographers used Cartesian coordinates--students taking response C on the pretest--to become disabused of that notion more successfully than students sho on the pretest thought geographers were using polar or celestial coordinates. The tendency, however, is very slight. The data in these cells are not large enough for me to make a very strong case for this interpretation.

Looking at the pretest data, we find that a number of students came into the course with the basic knowledge required to answer this question correctly: 1038 of them. The errors on the pretest tended to be distributed this way: 68 chose B, 92 chose C, and 38 chose D. Looking at the posttest data, the most interesting thing is of course, what happened in A, the correct response: there is a big decrease. But the

shift in B from 68 on the pretest to 92 on the posttest and the shift in C from 92 on the pretest to 287 on the posttest are also worth considering. The instructional process not only enhanced the relative popularity of B and C, but also the absolute popularity of these options. Thus it seems that students who entered the course with misconception C were more successfully moved to the correct response than were students with other misconceptions. Something happened, however, that disabused students who had the right notion at the beginning of the treatment -- that rid them of their rightness -- and introduced them to a misconception.

I again went back and reviewed the course description to see what I could find out about instruction on polar and Cartesian systems. It said these were introduced and dwelt on at some length. I came to the conclusion that the students had been saturated with these coordinates -- that more attention had been paid to them than to latitude-longitude -- and the result was that when they came to this question, they remembered having weeks of Cartesian coordinates and little on latitude-longitude. Hence they were sure that geographers must use Cartesian coordinates because they were given so much attention, and promptly responded "Cartesian."

This is my hypotheses -- a hypothesis formulated by someone who has never taught geography. However, even though I wasn't in the course and even though I haven't taught geography, inspecting the item results and reviewing the course description enabled me to set up a plausible hypothesis which I could discuss with the teacher.

The kinds of data we have been discussing here must be communicated to the teacher if he is to be aware of the effect of his instruction and if he is to alter his teaching, where necessary, to achieve the outcomes for which he is striving. If you present him with only the numerical data, however, he may become bored or frightened, particularly if he does not have the background to understand its significance. Put it in the context of what his students are learning, however, and you will find him interested in what you have to say. Also, if we could offer teachers this type of item analysis more regularly and routinely, they would soon develop sophistication in interpreting the results.

The last item I want to comment on is number 92 in Figure 3. This item shows the same pattern as item 98, the first item we discussed, in that there is, in addition to an increase in the number of students giving the right answer, a shift in the relative popularity of the distracters. The item is concerned with concept definition: "Which of the following must be included in a statement of a geographic fact?" The correct answer is C, "I, II, and III only." On the pretest, 487 students chose distracter A, 166 chose distracter B, and 556 chose D. Of the 487 students who began the course believing A, 250 or 51% ended the course giving the correct answer. Of the 166 students who initially responded B, 70 or 42% ended up correctly, while for D, the final result was 251 or 45% success. Thus, the results of instruction effectiveness varied very little for the three groups.

92. Which of the following must be included in a statement of a geographic fact?

I. Place
II. Time
III. Phenomenon
IV. Quantity

(A) I only
(B) III only
(C) I, II, and III only
(D) I, II, III, and IV

92.

Posttest \ Pretest	O	A	B	C	D	Total
O	4	5	1	3	11	24
A	3	46	19	26	21	115
B	4	42	34	32	35	147
C	14	250	70	193	251	778
D	16	144	42	118	238	558
Total	41	487	166	372	556	1622

Posttest \ Pretest	O	R	W	
O	4	3	17	
R	14	193	571	778 or 48%
W	23	176	621	
		372 or 19%		

Figure 3.

On the pretest, there were 1209 wrong answers and of these, 487 or 40% were A; 166 or 14% were B; and 556 or 46% were D. On the posttest, there were 820 wrong answers, and the results for the three wrong answers were A - 14%, B - 18%, and D - 68%. Thus, in terms of relative popularity as a mislead, distracter A was dramatically reduced from 40% to 14%, while distracter D rose from 46% to 68%. This is a latent effect of the instruction. D is wrong, as is A, but D is less wrong than A, and one result of instruction was to move responses from more wrong choices to less wrong choices. If we focus only on absolute correctness, we miss this type of outcome.

What I am urging with all these numbers is a more detailed examination of the patterns of responses to items on pretests given before a course and on posttests given after a course. The results of an educational treatment must be examined in terms of how the treatment has moved students from their initial differential positions to their final positions, if we are to evaluate its effectiveness. In the traditional treatment of achievement test results, you ask 50 questions and you accept a score of 45, without asking which 5 items were answered incorrectly. This use of results has its place in large-scale testing to determine over-all educational achievement status. When you come to curriculum evaluation, however, you have an opportunity to open the door to further information about the effects of instruction.

In particular, I am calling attention to shifts in the nature of error, or of misinformation, as outcomes of education. A student who confuses polar coordinates with Cartesian coordinates has displayed confusion but not ignorance, and teachers who center too narrowly on item success may miss important indicators of the further outcomes of education: partial outcomes, perhaps, and hence unsatisfying, but often partial journeys toward the goal.

Finally, I am advocating for curriculum evaluation the use of materials which might well be inappropriate for individual differentiation. Let us give more items at "time zero" - at the beginning of the course regardless of the fear that the group will

not know them. Frequently students will demonstrate a variety of misconceptions in their errors, and we can profit from this information. Education, in a sense, is a journey between points or positions. Let us measure these positions carefully, differentiating among errors, and let us assess the positions both before and after instruction.

ILLUSTRATIONS OF CURRICULUM EVALUATION PROJECTS

THE ROLE OF LARGE SCALE PROJECTS IN CURRICULUM EVALUATION WITH EXAMPLES FROM THE NATIONAL LONGITUDINAL STUDY OF MATHEMATICAL ABILITIES

Leonard S. Cahen
Research Psychologist
Educational Testing Service

My discussion today will cover two areas or topics. The first topic will discuss the role of large-scale curriculum projects in curriculum evaluation, with the National Longitudinal Study of Mathematical Abilities serving as an example. The second topic will be a quick survey of some of the issues and problems we face today in our responsibilities for evaluating educational programs.

The National Longitudinal Study of Mathematical Abilities (NLSMA) is a five-year study of pupil performance in mathematics.[1] An interim report on this project can be found in Cahen (1965).

The study was started in September of 1962 with three large testing populations at grades four, seven, and ten. The total testing population consisted of approximately 112,000 students from 40 states. Approximately 260 testing centers (school districts or a complex of schools) participated in the Project. The term "testing population" rather than

---

[1]The National Longitudinal Study of Mathematical Abilities is a research activity of the School Mathematics Study Group, Stanford University. The funding for this project came from the National Science Foundation. The principal investigator of the project is Professor E. G. Begle.

A series of Technical Reports on the National Longitudinal Study of Mathematical Abilities have been published. These Reports include reproductions of the test batteries used in the Study and the psychometric properties of the items and scales. Information about the availability of these Reports and how to purchase copies of them can be obtained from the School Mathematics Study Group, Stanford University, Stanford, Calif.

The comments in this paper reflect the opinions of the author from his former position as Project Coordinator and do not, necessarily, reflect the position or opinions of the project.

The figures shown in the presentation were supplied by Dr. James W. Wilson who was the Project Coordinator for the Analysis Phase of the Longitudinal Study.

"sample" is used, as the groups of participating schools cannot be considered a representative or a stratified population of schools or students. Participation in the study was voluntary. Bias, as typically reflected by higher levels of pupil aptitude in experimental curriculum programs, than for students in conventional curriculum programs, is a common observation in curriculum studies where participation in curriculum research is voluntary. NLSMA itself displays this aptitude bias which makes simple comparisons of differential effects of curriculum on achievement extremely tenuous.

Figure 1 shows the five-year testing schema for the three populations defined earlier.

The X, Y, and Z populations are the fourth-, seventh-, and tenth-grade testing populations respectively. The schema in Figure 1 shows that the same groups of X and Y students were tested for five years. The Z-population students were tested for three years with follow-up studies being made on these students after they left high school.

The students were tested each fall and spring. The test batteries included measures of mathematical performance, attitudes towards mathematics, anxiety, role and preference inventories, verbal and non verbal abilities, reasoning, cognitive styles, spatial visualization, and numerical facility. Some of these psychological measures were repeated at intervals throughout the study. Some of the mathematical scales and items were also repeated, but it was felt that the testing time could be used more efficiently in measuring specific mathematics achievemnt in depth rather than devoting a great deal of time to repeated testings with the mathematical scales administered in the earlier years of the study. The measures in depth included mathematical topics that are emphasized in specific school years in the curriculum, such as algebra and geometry.

The data will be used for assessing the long-term effects of different mathematics curricula on achievement, for learning more about the nature of mathematical abilities, for determining the relationship of early mathematics performance and later performance, for probing the complex problem of assessing change over time in mathematics performance, and for studying the correlates of pupil change in mathematics performance.

Figure 1 also shows the potential cross-sectional analyses of interest to the mathematicians who directed the Project. The reader will note that diagonal

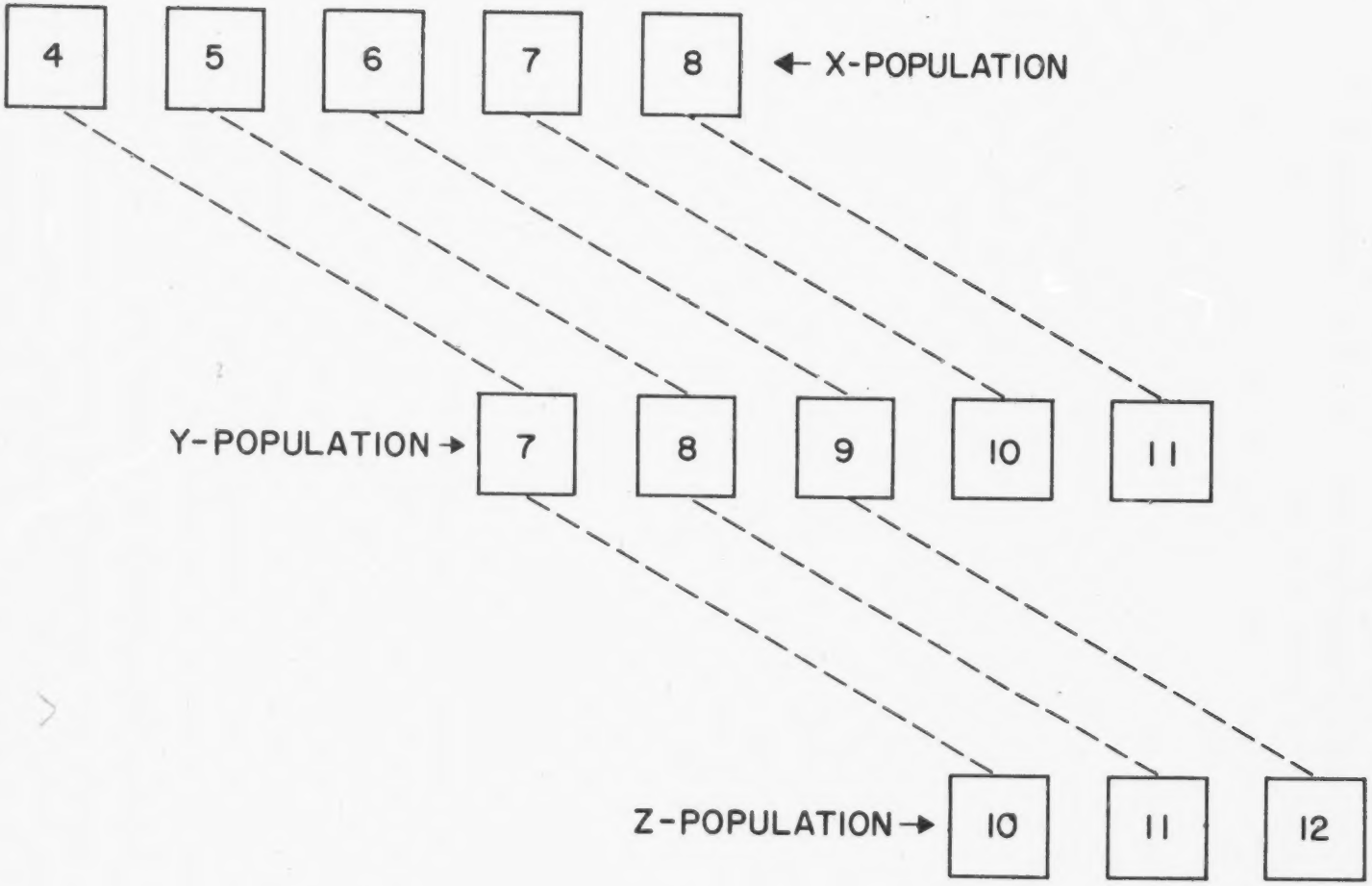


Figure I. Design of the Study

lines have been drawn from the X-population boxes schema in Figure 1 to the Y-population boxes, and similar lines have been drawn from the Y-population boxes to the Z-population boxes.

It has been hypothesized that X-population students who had modern mathematics in grades four, five, and six will perform at a higher level on NLSMA tests as seventh-grade students than did the original seventh-grade population.

The test of this hypothesis is outlined by the diagonal line from the fourth-grade X-population box to the seventh-grade Y-population box. The hypothesis stated above is grounded on the fact that few Y-population students had modern mathematics in their elementary school learning experiences, while many of the X-population students had this experience because modern mathematics curriculum materials were available in many of the schools these students attended in grades four, five, and six.

The test of this hypothesis makes the assumption that the X- and Y-population students being compared were comparable in aptitude levels and other crucial variables. Without a check of this assumption, no clear conclusions can be reached.

Figure 2 provides more detail about the extensive categories of data that have been collected on the NLSMA pupil populations, schools, communities, and teachers. The achievement and psychological batteries were briefly described earlier.

The information on teacher background and opinion, plus school and community descriptions, will serve two purposes. The information will be used to see if there is a statistical association between these measures and pupil achievement and attitudes. The information will also be used to describe, in statistical terms, the population of NLSMA teachers and schools to see whether these dimensions indicate bias favoring teachers and schools participating in one curriculum as opposed to another.

Figure 3 shows one type of analysis presently underway.

Schools that had a consistent curriculum sequence over the first three years of the study have been selected for the first analyses. By "consistent" we mean that the students in these schools were exposed to

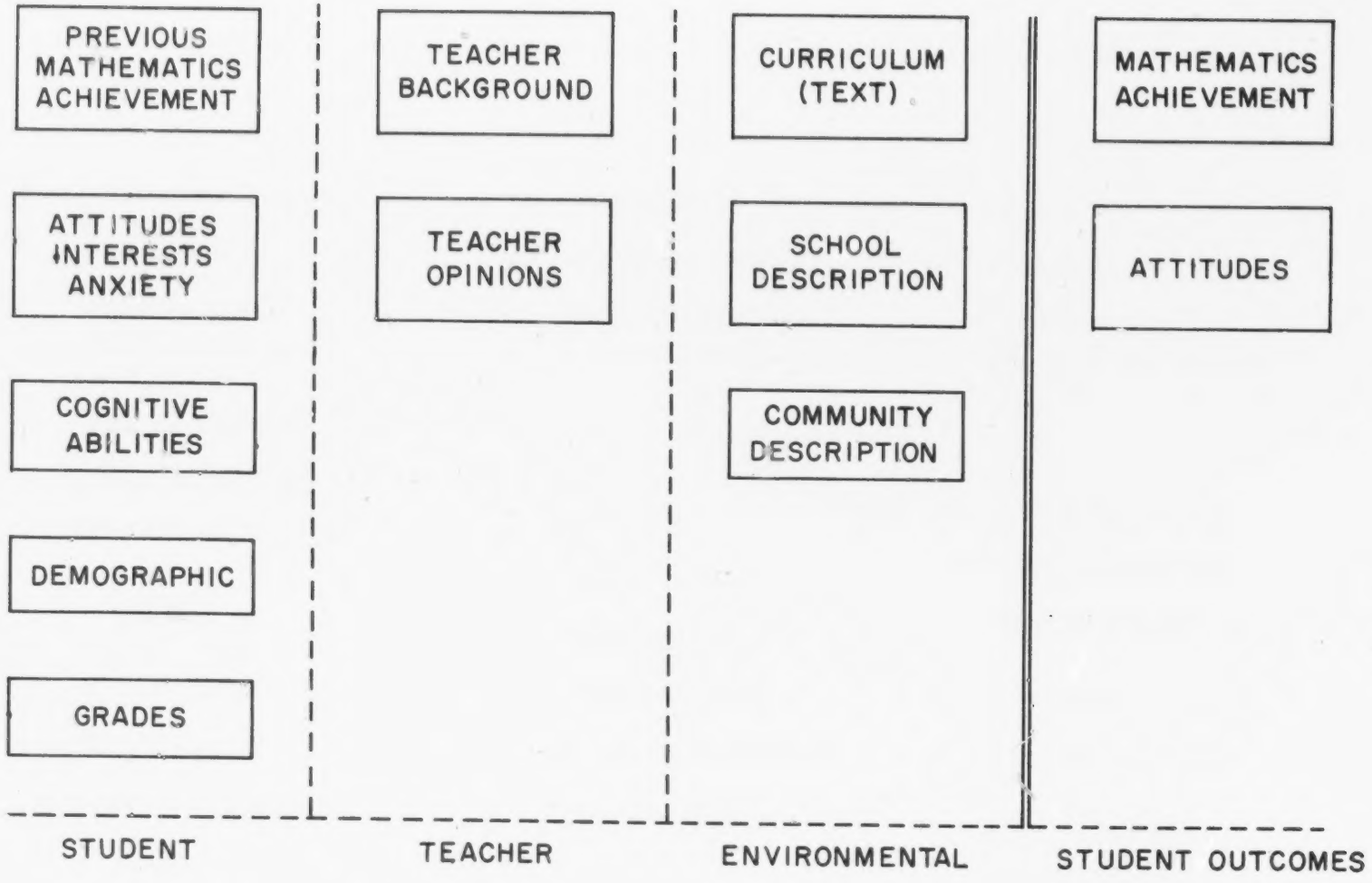


Figure 2. Classes of Variables Measured in NLSMA

GROUPS IDENTIFIED BY:	COMPARED ON:	WHEN INITIAL MEASURES ARE CONSIDERED:	
TEXT USED IN A GIVEN YEAR OR SEQUENCE OF TEXTS USED IN A GIVEN SET OF YEARS	SET OF MATHEMATICS ACHIEVEMENT MEASURES EACH YEAR	VERBAL AND NON-VERBAL ABILITY INITIAL LEVEL OF MATHEMATICS ACHIEVEMENT	
Independent variable is curriculum (or text)	Multivariate (several dependent measures)	Analysis of covariance	

Figure 3. Analysis to Identify Patterns of Curriculum Influence on Mathematics Achievement

curriculum materials in mathematics which were developed by the same publisher. In one sense, the mathematical materials may be described as forming a sequence of learning experiences for these students. Multivariate achievement measures in mathematics (school means) serve as the dependent variables. Analysis of convariance is to be used in an effort to "statistically equate" initial differences in aptitudes and performance across the various curricula which serve as the experimental treatments in classical terminology.[2]

The multivariate measures of mathematical performance (the dependent variables) will be looked at as a profile of scores. One curriculum may display a high set of means in one area (say, in computation) and a lower set of means in a second area of performance (say, understanding of concepts). This type of multivariate analysis breaks away from a single score reflecting outcome performance. A single score may mask very important differences in performance across treatment groups.

Other analyses are presently underway. These include differential performance within schools on the dimension of sex, differential performance within schools at different levels of pupil aptitude, and the development of models to assess change in pupil performance when some of the criterion measures of mathematics performance are repeated measures and others administered only once.

---

[2]The utilization of analysis of covariance was discussed in an earlier paper by George Temp. This technique was developed for the purpose of obtaining increased precision from the experimental design. This technique requires that the sampling units be randomly assigned to the treatments and was not developed for the purpose of equating pre-existing natural groups. Analysis of covariance has been used frequently in educational research and in curriculum comparison studies for the purpose of equating groups that differ initially on the convariates or "control" measures. The utilization of covariance for this special problem is a very difficult and complex one. The curriculum researcher should remember that randomization of experimental units to treatments is a necessary stage in analysis of variance and covariance models. Randomization allows the curriculum researcher to make much stronger inferential statements concerning the relationship of the treatment to results. The reader may wish to consult articles by Evans and Anastasio (1968) and Elashoff (1969) for general discussions about the utilization of covariance analysis. Frederic M. Lord (1967, 1969) has discussed the paradox in the interpretation of group comparisons when random assignment of sampling units to treatment groups is difficult, if not impossible, to obtain. An additional discussion of the covariance analysis paradox will appear in a forthcoming article by Charles E. Werts and Robert L. Linn (in press).

Before leaving this brief coverage of NLSMA, it is important to provide some information on the nature of the mathematics tests used in the study. A description of the procedures used to build NLSMA tests is to be found in an article by Romberg and Wilson (1968).

The mathematicians working with NLSMA felt it was important to examine performance in many different areas of mathematics. These areas included computation, understanding of number systems, algebra, geometry, problem solving, the ability to solve a new piece of mathematics, etc. As pointed out earlier, many of the items and scales from the NLSMA mathematics tests were repeated over the five-year testing program. Because of the limited testing time available in the schools, it was not possible or desirable to administer all scales or items to every pupil at each testing session. The item-sampling technique (Lord 1962, and Lord and Novick 1968) was used to supplement regular testing sessions where every student received every test item. In the item-sampling administration, a complete set of items is randomly assigned to different booklets and then the testing booklets are randomly assigned to pupils in the classes. For example, on one occasion we started with a 50-item test and randomly assigned five items to ten different forms or booklets. The items were mixed with other testing items. We used this technique to estimate school means on the 50-item test. Cahen, Romberg, and Zwirner (1970) report that the estimated school means obtained from the item-sampling procedure correlated in the high .80's with the estimated school means from the regular testing session where every student took each of the 50 items in a parallel form of the test.

Standardized tests were used by NLSMA on occasion. However, these tests typically yield one total score. Our mathematicians decided that items and scales drawn to their specific testing grid would be more efficient for the questions they were intending to answer.

The problem of creating testing instruments for curriculum assessment is sometimes a very difficult task. When one is evaluating a new curriculum effort, one rarely finds enough appropriate items in existing standardized tests. This is not to say that curriculum evaluation should avoid standardized tests. Items should be written, however, so that they reflect the specific goals and outcomes of the curriculum effort being evaluated. For a project like NLSMA, where many different curriculum

materials are represented, writing items that are "curriculum fair" is a difficult task. Care must be taken to see that the specific language and notation in the items do not favor one curriculum over another. Ideal items would tap areas such as the pupil's ability to perform mathematical problems that he has not seen before, and would therefore measure his ability to transfer his knowledge and skills to a new mathematical situation. These items are very difficult to write. Almost all of the items created by the mathematicians for NLSMA were pilot-tested, and items that appeared to favor one curriculum over another were modified or removed.

People responsible for curriculum decisions in school districts will require evaluative information from many different sources. Large-scale projects, such as the National Longitudinal Study of Mathematical Abilities, will provide one such type of information. On the other hand, it is important that evaluation studies be run at the local-school-district level in order to supplement the information obtained from large-scale testing projects such as NLSMA. The reasons for local evaluation studies are manyfold. A local evaluation project should be able to focus on topics and problems that are of primary and immediate importance to the local schools and its pupils. Information can be fed back much more quickly when the project is under the direction of the local school district.

One of the problems the NLSMA project had to face was the almost complete lack of feedback to the participating schools during the course of the study. Many of the schools are relying on NLSMA to provide them with evaluative information that they should have been gathering themselves. NLSMA can be looked upon as a form of summative evaluation in the terminology of Scriven (1967), yet earlier evaluation feedback information should have been available to the schools, either from their own internal research projects or from the various curriculum innovators who developed materials in the post-Sputnik era. On the other hand, a large-scale project such as NLSMA, has the potential of gathering data from a large population of students and therefore, it will be possible to generalize the findings hopefully, to students in other schools in the United States. The power of the generalizability of the findings is, of course, a function of the representiveness of students and schools in the study.

Another major advantage to a large-scale project such as NLSMA, is the fact that many extremely competent mathematicians and mathematics educators were brought together by the Project to develop and refine testing instruments. From a logical analysis of the items, it appears that many important dimensions of mathematical performance were covered in the NLSMA tests, and this type of

instrumentation is difficult to develop in small-scale local projects. Dr. Hulda Grobman (1968) has prepared a monograph which deals with many aspects of evaluation in curriculum projects. The topics and areas covered in the monograph will be helpful to those concerned with evaluation -- whether the effort be on a large or small scale.

The last section of my presentation will be devoted to a quick survey of some current evaluation problems. These problems face the evaluator, whether he is working on a small-scale local project or on a large-scale project, such as the Longitudinal Study. If the assumption is made that it is important and necessary to gather outcome performance measures in many areas rather than letting a single score reflect the level of output, research is needed to determine how decisions are reached from a wide array of performance measures -- some of which may be positive and supportive of the curriculum effort while others may be negative. We need to know a great deal more about the range and types of measures we need in order to reach decisions about curricula. Too often evaluation efforts gather only information that is immediately relevant to the specific curriculum. Schools need to assess the possibility of negative side effects as well as the short-term effects of curriculum innovation.

A second issue is the one concerning the logic and utility of curriculum comparisons. Let us assume we have two different curricula which we will call A and B. Let us also say that curriculum A is an experimental curriculum, while curriculum B reflects a curriculum that has been in existence for a number of years. On the surface it would appear to be illogical to compare the outputs of curriculum A and curriculum B because they were designed to yield quite different types of pupil behavior. Michael Scriven (1967) has suggested, however, that information should be made available on many outcome dimensions, so that the consumers of information at the school-district level can make a decision which takes into account positive as well as negative outcomes on a wide range of pupil performances across the competing curricula. The consumer must let his own value system play an important role in reaching a decision as to which curriculum he should adopt for specific learning situations in his school district.

An alternative to curriculum comparison studies is to select a curriculum on logical grounds and then direct the innovative and evaluative

efforts toward the process of learning how to teach the materials effectively. This type of curriculum research would enable school districts to pursue the problem of how to make the materials teachable to students with a wide range of aptitudes and backgrounds, and would also provide important information about how teacher variables might interact with the curriculum materials and differential pupil backgrounds and aptitudes. The technique of devoting the evaluation efforts to a single curriculum, rather than comparing curriculum A versus curriculum B, alleviates many of the problems of making comparisions across different curricula where there are systematic biases favoring students in one curriculum versus a competing curriculum.

Before closing, I would like to suggest that people in curriculum research take advantage of the important information that is contained at the item level as well as looking at total performance over a set of items. Item analyses can provide evaluators with a great deal of information about the types of errors made by pupils. This type of information can be used efficiently in modifying curriculum materials, as errors may indicate where curriculum materials are not clear to pupils and provide information to teachers about the effectiveness of their own teaching.

REFERENCES

Cahen, L. S. An interim report on the National Longitudinal Study of Mathematical Abilities. Mathematics Teacher, 1965, 58, 522-526.

Cahen, L. S., Romberg, T. A., & Zwirner, W. The estimation of mean achievement scores for schools by the item-sampling technique. Research Bulletin 68-39. Princeton, N. J.: Educational Testing Service. To appear in Educational and Psychological Measurement, 1970.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-401.

Evans, S. H., & Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. Psychological Bulletin, 1968, 69, 225-234.

Grobman, H. Evaluation activities of curriculum projects: A starting point. AERA Monograph Series on Curriculum Evaluation, Vol. 2. Chicago: Rand-McNally, 1968.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Lord, F. M. Statistical adjustments when comparing preexisting groups. To appear in Psychological Bulletin, 1969, 71.

Lord, F. M., & Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Romberg, T. A., & Wilson, J. W. The development of mathematics achievement tests for the national longitudinal study of mathematical abilities. Mathematics Teacher, 61, 1968, 489-495.

Scriven, M. The methodology of evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, Chicago: Rand McNally Co., 1967.

Werts, C. E., & Linn, R. L. Lord's paradox: A generic problem. To appear in Psychological Bulletin.

IN-SERVICE EDUCATION PROGRAMS

Thomas S. Barrows
Associate Research Psychologist
Educational Testing Service

Teacher-training programs, like all others, present problems to the evaluator. A number of them are common to other evaluation settings, but one--the one I would like to discuss with you here--seems to me to be of special interest in the teacher-training context. It is simply, "Who, what, and where to measure?"

Consider for a moment that we are to evaluate a summer workshop consisting of phonics instruction for reading teachers. Shall we measure teachers' knowledge of phonics, observe how much phonics they teach in their classrooms, or test their students' knowledge of phonics?

A strict theorist will suggest that adequately specified behavioral objectives will determine that decision. On the other hand, most of us know, or at least suspect, that the evaluator's real world is not like that. Put optimistically, the specification of objectives and the planning of evaluation combine in an interactive process. Or, if you prefer, the evaluation specialist usually ends up specifying objectives because others who are supposedly responsible for this task are unable or unwilling to do so. My experience has been that the evaluation specialist at least enters into the process. How, then, is he to select among the "who, what, and where" options?

I believe that the evaluation specialist should consider four factors when refining or recasting objectives and planning an evaluation. These are decision relevance, design constraints, data-collection techniques, and cost. He should ask, "What information will govern decisions to revise, continue, or discontinue the program?" At what point is design the strongest? At what point is measurement the strongest? What evaluation costs will the project bear?" His ultimate decision of who, what, and where to measure will be governed by a subjective weighting of the responses to these questions.

Let me outline for you three teacher-training projects that ETS has evaluated or is currently working on. The characteristics of these projects and the choices of evaluative techniques should serve as examples of the above considerations.

Example One

This teacher-training program came to us with poorly defined objectives. The directors were sure that their project was intended to change teachers' behavior towards minority-group children in such a way that students' self-concepts would improve. In addition, teachers were to become more helpful and knowledgeable about solving the problems that the children and their families have in coping with school and society. The program's treatment consisted of prescribed reading, addresses by recognized experts (i.e., Pettigrew and Clark), and discussion periods.

We considered cost and measurement techniques first. A review of the literature indicated that valid measurement of behaviors included in the vague construct of self-concept would require costly instrument development. Furthermore, we could not be sure that we would ever be successful because of the poor definition of the self-concept and documented failures in previous attempts at instrument construction. Similarly, the task of measuring students' degree of freedom from problems in coping with school and society appeared formidable and expensive. Thus, the consideration of measurement and cost problems alone suggested looking at teachers.

Decision relevance also suggested teacher behavior. Consultation with the project's staff suggested that the program's effectiveness could be judged and decisions to modify, continue, or discontinue the program could be based on change in teachers' attitudes and change in teachers' knowledge of ethnic group characteristics and intergroup relations. We discussed the relative importance of each objective and reviewed the availability of instrumentation for each. We found nothing suitable in either area and judged the cost of constructing and validating an attitude instrument to be greater than similar costs for a factual test. We decided to try the factual area first.

Design presented no problems at the teacher level. Happily, there were more applicants than could be accommodated and we could assign applicants to treatment or control groups at random. Controls were told that they had been accepted for future cycles of the program in an attempt not to alienate them.

Finally, cost of the entire study of effectiveness in attaining these cognitive or knowledge objectives was within the program's budgeted capability. We carried out the study, and completed our report in late June of 1967. Our findings were compellingly negative and, as might be expected, our proposal to look at teacher attitude change and teacher classroom-behavior change was not accepted.

The report of the study, (Operation Upgrade: Effectiveness in Attaining Cognitive Objectives. Project Report, Educational Testing Service, 1967), is available on request, so I will not go into further detail. It seems more profitable now to consider Example Two.

Example Two

The second example concerns a summer program for about 85 kindergarten-through-2nd-grade reading teachers. The workshop will last five weeks and the primary objective is to acquaint the teachers with a number of divergent reading programs so that they may tailor their instruction more adequately to students' individual needs. The treatment will consist of lectures, model lessons, and directed practice teaching.

Decision relevance has been our first concern. The directors of the workshop definitely feel that it would be unrealistic to hope for improved student achievement in the first year following the workshop. They also point out that teachers' knowledge of reading programs does not insure use of the programs. Thus, they minimize the value of data gathered on teachers' knowledge resulting from the workshop. The directors favor observation of teachers' in-class behavior as the basis for their decisions.

From a measurement point of view, we would very much like to use either teachers' knowledge following the workshop or student achievement. For the latter, we have excellent instruments available, while the former would require only some comparatively simple achievement test construction. On the other hand, teachers' in-class behavior will require a fairly sophisticated time-sampling technique, the training of observers, and the construction of an observation schedule. The establishment of interobserver reliability adds

complications, and additions must be made to the final analysis to account for observer variance. Furthermore, the reactivity of observations is well known. It is pretty clear that teachers do not give typical performances when observers are present.

The cost of obtaining observation data also mitigates against measurement of teachers' in-class behavior. In order to obtain the data on all 85 teachers at approximately the same time, we will need many observers who will have to be trained and paid. Paper-and-pencil instruments for students would clearly be cheaper, and the same is probably true of the construction of a test of teachers' knowledge.

Finally, we must consider design. It is not yet clear that there will be more applicants than can be admitted as participants in the program and so an experimental/control design with teachers or their students as subjects cannot be chosen with assurance. If it is not possible to constitute random experimental and control groups, a pre/post design might be used. Such a design cannot be applied to student achievement data without accepting the highly questionable assumption that the participant teachers' 67-68 and 68-69 classes are initially equivalent. The naturally weak pre/post design is thus rendered even weaker by this situational constraint. It appears, then, that design considerations suggest dropping student achievement data and using teachers' knowledge or in-class behavior if true experimental and control groups are not available. Such a pre/post design would, of course, collect pretest data on one random half of the teachers and posttest data on the other random half in order to obviate possible practice effects.

In order to come to a decision as to who, what, and where to measure, it is now necessary to integrate these considerations:

1) Decision relevance suggests teachers' in-class behavior.
2) Design suggests either teachers' knowledge or in-class behavior.
3) Measurement suggests teachers' knowledge or student achievement.
4) Cost suggests teachers' knowledge or student achievement.

We chose teachers' in-class behavior as the data to obtain. Clearly, our subjective weighting stressed decision relevance and design at the expense of measurement and cost considerations. A simple addition of considerations favoring each type of data would have indicated teachers' knowledge 3 to 2.

Example Three

Initially, this project involved curriculum construction of grand proportions. A kindergarten-through-12th-grade social studies curriculum was to be constructed, emphasizing problem-solving skills and the synthesis of concepts from history, sociology, economics, and anthropology. Although the proposal was not entirely explicit, we assumed that course outlines, lesson plans, student materials, and all the other trappings of a total curriculum would either be produced or selected from existing sources. Teacher training was included explicitly in the form of both a summer workshop and additional released time during the normal school year.

In view of the apparent comprehensiveness of what we thought was a total curriculum effort, our initial plans for evaluation were quite exhaustive. We planned to use the following types of data: teacher ability and attitude, teacher in-class behavior, student in-class behavior, and student achievement and attitude. This exhaustiveness seemed to be called for because specific aspects of the program could be expected to have their impact at different points. It seemed desirable to be able to isolate the effectiveness of separate aspects of the program. In addition, the program's financial capability allowed us complete freedom. While the dollar amount budgeted for evaluation was not great, the project's staff agreed to undertake a large share of the item-writing responsibilities in cognitive areas. In this best of all possible financial worlds, neither measurement nor design constraints suggested deletion of any of the types of data. Design possibilities were equivalent at all levels, and there was no need to economize on the number of variables collected.

All of this was too good to be true. Things have changed as the program has progressed. Curriculum construction is sadly behind

schedule. (I suspect that this is true in a large number of Title III projects.) Goals, as a result, have been quietly revised to focus more heavily on the teacher-training function. The project staff now intends to conduct the originally planned training sessions and to furnish teachers with model lessons which should further influence teaching techniques and content. From the point of view of evaluation, we are now looking at a teacher-training program.

We feel now that there is a cost restraint. It is the extent of expenditure which we can personally justify in view of the project's slow progress. It would seem both wasteful and improper to spend large sums of money evaluating a program for which the expected educational payoff is small.

Decision relevance originally suggested that we look at all three levels of data--teacher characteristics immediately following the workshop, teachers' in-class behavior, and student attitude and achievement. The project's poor progress now suggests that expecting change in student attitude and achievement is unrealistic and that decisions regarding continuation, suspension, or revision of the project will be made on the basis of both teacher characteristics following training and teacher in-class behavior.

The designs which are applicable are still identical at all levels. Random assignment is not possible with either teachers or students. Non-equivalent comparison groups of both teachers and students are, however, available.

Measurement strengths and weaknesses have received much of our attention. At the level of teacher characteristics, we had proven attitude measurement techniques which were adaptable for our purposes. Given the project's vague objectives, we reasoned that attitudes toward sociology, economics, history, and anthropology, should be improved if teachers were to teach in a multidisciplinary fashion. We also found an existing instrument which seemed appropriate for teachers, and which purported to measure their ability to formulate hypotheses to explain given data. We felt that this might operationalize the vague problem-solving objective.

Teachers' in-class behaviors, on the other hand, would require construction of an observation schedule, the training of observers, and all the other complications mentioned in Example Two above. Student achievement data

would require relatively simple instrument construction, and an appropriate instrument for student attitudes was found. Thus, these measurement considerations suggested looking at teacher characteristics and pupil achievement and attitude.

What did we decide to do in this project? We are still deciding as the nature of the project changes, and the resulting effects on our four considerations change. We can, however, summarize current status in this on-going decision process:

1) Decision relevance suggests teachers' knowledge and attitude following training and their in-class behavior.
2) Design considerations do not enter into the decision.
3) Measurement suggests student achievement and attitude, and teachers' knowledge and attitude following training.
4) Cost suggests student achievement and attitude, and teachers' knowledge and attitude following training.

It appears as though decision relevance and justifiable cost will be heavily weighted in our last analysis, and we will therefore probably limit our evaluation to teacher characteristics.

Finally, let me review my examples. In the first, cost, design, measurement, and decision relevance seemed to point to one type of data. In the second, a subjective weighting was necessary to resolve contradictions between considerations. In the last example, decision relevance and cost generally seem to be dictating the decision, although a final one has not been made.

I hope that these three examples indicate to you the variation that can be expected when considering several teacher-training projects. A greater hope is that the examples have somehow suggested that there is no one best type of data which should be collected in every project, but that a careful consideration of cost, design, measurement, and decision relevance is always necessary for a unique, appropriate decision.

THE UNGRADED PRIMARY SCHOOL

(Chesapeake, Virginia Public Schools)

J. Robert Cleary
Director, Field Services
Educational Testing Service

We thought that the project that I will describe briefly would be of interest to you for several reasons:

1) Its scale is different; that is, a Primary School is the unit of study--not a region, a curriculum project involving many areas, nor even a school system.
2) It illustrates an Educational Testing Service response to needs in the field via consulting, as contrasted with a project or a research study response.
3) It is current--in fact, it is really in its beginning stages.
4) It illustrates that a good deal can be accomplished by a school staff with only modest support from outside sources.

Following is a brief description of the community and the background of the project.

Chesapeake, Virginia, is a region to the south of Norfolk which previously had been Norfolk County. Several years ago, the community was incorporated as a city, although it has no real resemblance to a city: it has wide reaches of woodland, and extensive rural areas, as well as small clusters of suburbia composed of middle and lower middle-class dwellings. As a result of incorporation, sections of Portsmouth and Virginia Beach are now included in Chesapeake, Virginia. It has 25 elementary schools, 6 high schools, and a new vocational-technical center. It serves a school population of over 26,000 students drawn from a geographical area, referred to as a city, which is one of the largest in the country.

In November of 1966, a proposal to erect a structure, obtain staff, and develop a program for an ungraded primary school to serve five-, six-, and seven-year-olds in the community was submitted under Title III of ESEA. (Eight-year-olds were to be added the following year.) The proposal was funded in the spring of 1967, but expenditure of funds was delayed until the summer of 1967.

By the middle of September of 1967, a temporary structure with facilities and flexible spacing arrangements had been erected, staff had been recruited, and some instructional materials were on the scene, when some 360 five-, six-, and seven-year-olds descended on the school. What kind of school? What kind of program?

In the interest of brevity, but at the risk of over-simplication, I shall characterize the main elements of the Ungraded Primary School Project as follows:

1) Its view of the child is a developmental one -- developmental in the sense that it believes the child will progress more quickly, more easily, and yet more efficiently, if direct efforts are made to provide the next logical increment in the continuum of progression in all areas of school life. This might be contrasted with the extreme of the "readiness" view which might say, "He is not ready; let him mature a year before providing a formal experience of some kind." The pupil in Chesapeake who cannot skip or cannot walk a balance beam is provided with opportunities to skip and to balance himself on a beam. The pupil who cannot make certain visual discriminations is "taught" to make them, but with no pressure.

2) Its form is ungraded, because staff know that neither age groups nor grade groups can accommodate as well the variations in the present achievements of the students or in their rates of learning.

3) Like all schools, it must work with pupil groups -- but groups, no matter how temporary, should be composed of pupils requiring the same or similar educational experiences at that time, regardless of their age or their levels of achievement in other parts of the school program.

4) Its instructional strategy is to use all the information available about the pupil and to "lead to his strength -- patch up his weakness," recognizing also that different learners have different learning modes.

5) It attempts to use the latest information contributed by education, psychology, and other disciplines in designing and revising its program.

6) It employs flexible scheduling, flexible space arrangements, flexible staffing patterns, and variations in materials of instruction and learning

conditions to match learning experiences to the learner more precisely.

Even the bare sketch just presented suggests the need for a great deal and many kinds of pupil information. In addition, the fact that the school was new and that there was no formal kindergarten program in the schools of Virginia meant that the school staff had no information on approximately two-thirds of its pupils -- the five- and six-year-olds.

Thus, the staff was faced with the problem of designing and organizing an assessment program utilizing instruments sensitive enough to provide the differential information required to support and evaluate the program. We began our discussions with the staff toward this end in the middle of the summer of 1967. By the end of September, we had assembled a pre-assessment battery of over thirty components -- task sets, locally prepared tests, and standardized tests. Informally, we have adopted this nomenclature because of the differences in length and function of the instruments.

These instruments provide pupil information in six broad areas:

1) Verbal characteristics - including formal reading and spelling skills, listening skills, oral language, auditory discrimination, visual discrimination, letter knowledge, and visual-motor skills.

2) Mathematics characteristics - including formal mathematics understanding skills, mathematics fundamentals, and more basic mathematics understandings.

3) General intellectual development.

4) Specific instructional information, such as letter and numeral knowledge, phonics knowledge, and specific mathematics and science understandings.

5) Physical characteristics.

6) Social-emotional characteristics.

Over half of the instruments were developed locally by the staff with some assistance, after elaborate inspection of available instruments failed to uncover suitable sampling of objectives. The standardized tests which were used survived careful screening and selection.

Testing began in October of 1967 and proceeded normally, complementing the natural get-acquainted instructional activities in the classrooms which were planned as self-contained for the first few weeks. The self-contained procedure was employed as a technique to provide an easier transition to the school and to allow time for the collection of differential information from the assessment program, which was used as a basis for initial placement in groups in the ungraded program.

The pre-assessment battery was developed, then, for the following purposes:

1) To obtain necessary pupil information.

2) To provide base lines for project reporting and evaluation.

3) To provide differential information for the initial scheduling of students.

I shall not take time to describe the procedures for scheduling pupils, except to say that not all of the pupils were tested with all of the instruments nor were the results on all instruments used for scheduling.

The pupil record card is the product of the work of the project staff. We developed a key battery to administer to all pupils. Then we developed a procedure for reporting, which directed pupils upward from the key battery for more information on higher levels of achievement or downward for more information, in the case of chance-level performance on measures in the key battery. In this way we were able to stage the testing program to obtain the information required, without administering all instruments to all pupils.

This procedure worked very well, although we must admit that the time and organizational detail necessary to pull it off was more than we had bargained for. However, the staff reacted positively to this technique, and teacher validation of the initial selection and assignment to groups was unbelievably high. Only four pupils were reported by teachers as misplaced seriously enough in initial learning experiences to warrant changes in schedule or assignments.

I shall take a moment now to describe the reporting procedure we developed for the instruments. A score from each cognitive measure, whether from task sets, local tests, or standardized tests, is pooled with all other pupil scores -- regardless of age -- to form one distribution of scores. Means and standard

deviations are calculated as well as standard errors of measurement. The mean chance score and the standard deviation of the chance scores are also calculated for each test in multiple-choice form. We adopted the symbols H, M, L, and this symbol - Ø - to indicate a region in the distribution within which the scores could have been obtained by chance. This chance region is established for each distribution by adding twice the standard deviation of the chance scores to the mean chance score. We can be almost certain (approximately 95 chances out of a hundred) that scores above that level could not have been obtained by chance and are therefore indicative of knowledge. The symbols H and L are derived by adding and subtracting one standard deviation from the mean test score of the group. M, the remaining middle region, lies between plus and minus one standard deviation above the mean. This, then, is our method of reporting what really are standard score regions. Point scores are not reported, and standard errors of measurement are always considered in favor of the higher score classification.

We use a short-cut formula for standard deviation, and the other formulas are quite simple. Three members of the project staff have now been trained so that they can henceforth perform these operations.

Thus far I have given a brief sketch of the community and how the project developed, described the pre-assessment battery, and indicated how results on most instruments are reported.

Next, I should like to demonstrate how the assessment battery supports the educational program. One example will serve to illustrate how information from somewhat sensitive instruments was coupled with a logical or functional analysis of certain materials of instruction, in order to evaluate pupil performance as a basis for initial placement in the reading sequence.

A first step was to ask the data three questions, the answers to which served to organize the pupils into three broad groups initially. The three questions were:

1) Who is reading? That is, who has broken the code?

2) Who is ready to begin?

3) Who needs more preparation before beginning to break the code?

For the reading and the "ready to begin" groups three reading approaches were analyzed: the Lippincott, a rigorous and rapid approach; Houghton-Mifflins' McKee, a somewhat less demanding approach; and Sullivan's programmed materials approach. Analyses of the requirements or the "press" of these materials led to the following specifications for information supplied by the assessment battery:

> For Lippincott success, a pupil would need high performance on all three related skill areas: listening, visual discrimination, and auditory discrimination.
>
> For McKee, a pupil would need middle performance in listening, high visual discrimination, and middle-level skill in auditory discrimination.
>
> Programmed reading seemed most appropriate for pupils with low listening skills but with middle-level auditory skills and middle or high visual discrimination, provided their copying scores were high enough to insure the fine motor coordination necessary to contend with the requirements of the workbook exercises.

The next step was to relate each component of the pre-assessment formally and permanently to sequences of learning experiences in some of the areas mentioned earlier. From the beginning, the curriculum sequence had been more implicit than explicit in the planning memoranda and in the proposal. Although this is still mainly true, some progress has been made in specifying curriculum elements and sequences.

From the outset, the project staff generally, although not wholly, subscribed to the notion that educational objectives, particularly for this school, must be stated in behavioral or operational terms. But with pressures of time, they have made only modest beginnings.

Although the assessment battery falls far short of representing samples of all educational objectives in a given curricular sequence, we decided to make progress along these lines by placing our instruments in sequence and translating what they measured into behavioral statements.

Thus we developed a pupil record which we presently call Individual Pupil Behavioral Characteristics. It is at once the beginning of a behavioral record system, a trace system for individual pupil progress, a record of teaching strategies and their results, and perhaps a useful way of reporting pupils' progress to parents. In short, we have the beginnings of a communication system using the only important message unit: what the pupil can do.

Eventually this working document will have many additional statements which will represent the major behaviors in the curricular sequences, many of which will not be measured by project instruments. The document is only a first draft of a beginning. It is, however, our first effort to remind all involved that evaluation is a part of the instructional process, not apart from it.

That is where we are in Chesapeake.

Let me highlight our next steps:

> We shall test again in the spring of 1968 (?) with fewer instruments. The first year was shakedown cruise. The test analyses and item analyses I have already performed show that a few instruments are not working well and that there is some redundancy in the battery.
>
> We shall analyze certain pretest-posttest data -- notably in Reading, Listening, Fundamental Operations, Understanding Mathematics, and Logical Reasoning, and we will use the behavioral record to test pupils at higher stages than those they occupied in the fall.
>
> We shall also analyze pupil rates of progress on the sequence of behavior indicated on the behavioral record in relation to their learning experiences whenever possible.
>
> We intend also to look for trends or patterns which might be isolated for future spin-off research studies.
>
> Finally, over the long haul, we shall begin to take a systems look at the entire enterprise without the benefit of hardware.

In short, we shall attempt to use whatever we know about measurement and evaluation to give practical support to the teachers' efforts to make teaching and learning more efficient.

It is too early to tell how successful this project will be, but it will bear watching. One thing is certain: they are moving, and they get things done.

CONCLUDING REMARKS

John S. Helmick
Vice President
Educational Testing Service

As you've noted, the last item on the program is "Concluding Remarks" by the Chairman. I'm glad it was listed as "Concluding Remarks," and not as "Summary," because I certainly don't intend to try to summarize what has gone on during the last day and a half. Nevertheless, I would like to conclude with a comment or two.

I think we have kept our promise that we wouldn't make experts out of you in this period of time. I believe we've also demonstrated that we do not have all the answers. I hope, however, that we all see some of the problems a little more clearly as a result of this discussion. I'm sure we see the problems more clearly than we see the solutions. Nevertheless, I feel the situation is far from hopeless, even though we can't, at this stage, come up with the simple cookbook how-to-do-it approach to curriculum evaluation.

We do have to make decisions about curricula. We do have to decide to continue a particular program or to discontinue it, to innovate or not to innovate. (As I think has been pointed out particularly by Mr. Barrows, maybe it's not really innovation: it's just renovation or reinnovation.) We have to make these decisions on the basis of the information we have, and the information we have should be the best available. If we can improve the information available and if we can improve our understanding of it, even though the information is not perfect, we can make better decisions affecting choice among curricula.

What we are forced to rely on, to a large extent, is common sense. The unfortunate thing is that common sense is very uncommon. Maybe "informed judgment" is a better term for what I have in mind. We're not going to come up in the near future with automatic techniques which tell us how to obtain a number that will give us the answer--if it is larger than a fixed value, we decide "yes;" if it's smaller, we decide "no." We are therefore going to have to use judgment, which should be as informed as possible.

It's my hope that by simply having considered some of the problems, some of the approaches, and some of the ramifications of curriculum evaluation, you are in a little better position to make informed judgments and to assist

others in making informed judgments. We've given some examples and we've described some techniques, but we certainly haven't produced any foolproof way of making the right decisions.

There is one approach I think is worth special mention, since it has come up in a variety of contexts. It seemed to me that a number of people were saying that in making decisions in the general area of curriculum development, we should not rely unduly on abstract total scores. We need to look at the behavior that is behind these scores, and at the ways in which that behavior is related to the ongoing educational process. This point was made in several of the discussions, and particularly in Tom Donlon's presentation on item analysis, or even the somewhat untraditional forms Tom described, which are related to rather simple multiple-choice type questions.

Another point that should be emphasized is that, even though we are concerned with the relationship of curricula to individuals, our approach to curriculum evaluation is different from our approach to individual evaluation, and the appropriate techniques for the other. That is, you don't do the same kind of item analysis interpretation for an individual response to a single item that you do when you are dealing with a group response to a particular item.

Well, we've enjoyed it. As I predicted in my introductory remarks, it has been useful for us to get the kind of interactions we've had with you on a variety of these topics. We've profited from the interchange. We only hope that you have too, and that you can now approach evaluation in a somewhat more understanding and effective way.

Thanks again.